ADAPTIVE VISUAL KV CACHE REUSE FOR EFFICIENT MULTI-TURN ITERATIVE MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-turn visual reasoning acquires evidence through repeated cropping and local inspection, but overlapping observations lead to redundant visual encoding and prefill. We propose an adaptive visual KV cache reuse framework that reduces repeated processing while retaining access to additional image detail. The central difficulty is that skipping visual prefill also removes the interaction between a new observation and the current reasoning: historical keys remain bound to old positions, while historical values have not incorporated the intent formed between crops. Our framework first selects full re-encoding or regional reuse according to crop size and cache availability. On the reuse path, Positional Key Rebinding (PKR) corrects attention matching at the new positions, and Intent-guided Value Enhancement (IVE) uses pre-crop reasoning attention to strengthen content relevant to the current objective. Source states from fully re-encoded crops are retained for subsequent retrieval and reuse. The framework requires no additional training; on reuse hits, it bypasses the visual encoder and prefill for the selected visual tokens, then resumes normal text reasoning. Experiments show that our method reduces inference time while preserving the overall accuracy of multi-turn visual reasoning that relies on intermediate crop images.

1 INTRODUCTION

Multimodal models increasingly acquire visual evidence during reasoning rather than interpret images once, using existing observations to locate, crop, and magnify regions for further inspection (Lai et al., 2026). V* uses LLM-guided visual search to locate fine-grained evidence, Visual Sketchpad integrates visual tools and their outputs, and ICoT interleaves image patches with textual rationales (Wu & Xie, 2024; Hu et al., 2024; Gao et al., 2025). DeepEyes and Pixel Reasoner learn active inspection through reinforcement learning, while Mini-o3 extends visual search to longer, more diverse trajectories (Zheng et al., 2026; Wang et al., 2025a; Lai et al., 2026). Together, these advances make perception an ongoing process of acquiring evidence as reasoning progresses.

Advances in model design and optimization improve performance (Ye et al., 2026; 2025a;b), but multi-turn visual reasoning also demands efficient inference. Standard pipelines encode each returned crop and prefill its visual tokens (Figure 1(a)), repeatedly processing overlapping regions without using their existing states. Visual token compression reduces computation within an observation (Chen et al., 2024; Zhang et al., 2025b; Yang et al., 2025; Shang et al., 2025), whereas multi-turn reasoning also permits KV reuse across observations. Since cropping and magnifying original pixels remains essential for new detail, we study how to reuse prior visual computation across successive crops while retaining full re-encoding for further inspection.

Such reuse requires historical visual states to fit the current reasoning context. Standard visual prefill constructs keys at their new insertion positions and lets visual representations interact with the preceding reasoning through successive layers. Direct reuse omits both computations (Figure 2). Historical keys retain rotary position embedding (RoPE) encodings for earlier positions, changing query–key matching and potentially disturbing attention allocation between visual evidence and reasoning text (Su et al., 2021). Even after positions are corrected, the values still come from an earlier observation: intervening reasoning may have shifted the objective from locating an object to reading text or identifying an attribute, but this new crop intent did not participate in constructing

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.

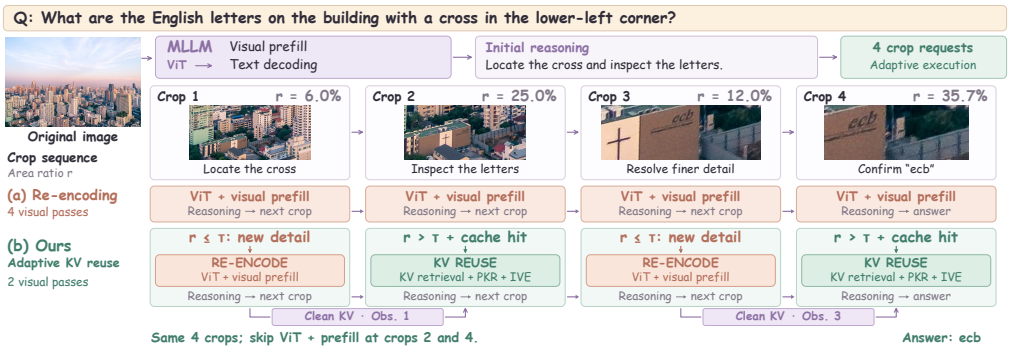


Figure 1: **Repeated re-encoding versus adaptive visual KV cache reuse.** For the same four crop requests, (a) full re-encoding repeats visual encoding and prefill at every step, whereas (b) adaptive routing re-encodes crops 1 and 3 for new detail and reuses their cached states for crops 2 and 4 through PKR and IVE, skipping visual encoding and prefill for two crops.

the cached values. Visual KV reuse must thus address positional mismatch and evolving crop intent so that existing visual content can support the next reasoning step.

Based on this analysis, we propose *adaptive visual KV cache reuse* (Figure 1(b)). The framework uses relative crop size and cache availability to choose between image re-encoding and regional KV retrieval. For reused regions, *Positional Key Rebinding* (PKR) applies rotary encoding to clean source keys at the current insertion positions, restoring their intended relative positions for attention. *Intent-guided Value Enhancement* (IVE) uses pre-crop reasoning attention to the parent observation to strengthen values relevant to the current objective. Together, they connect historical visual states to the new reasoning step, bypassing visual encoding and prefill for the selected tokens. The framework requires no additional training and leaves the model's crop-decision mechanism unchanged, although realized reasoning and crop trajectories may differ.

Our main contributions are as follows:

- **Adaptive visual KV cache reuse.** We select between regional reuse and full re-encoding based on crop size and cache availability, reducing repeated computation while retaining access to new visual detail.
- **Positional Key Rebinding.** We analyze positional mismatch in direct reuse and rebind cached keys to their current insertion positions, restoring the intended relative positions in attention.
- **Intent-guided Value Enhancement.** We use reasoning between crops to strengthen values relevant to the current objective, without rerunning prefill for the selected visual tokens.
- **Accuracy and efficiency.** Across five benchmarks, our method preserves or improves overall accuracy while achieving 1.30–$1.40\times$ end-to-end speedups.

2 Related Work

2.1 Iterative Visual Reasoning

Iterative visual reasoning lets models inspect an image further based on intermediate reasoning. V* uses language-guided search to locate fine-grained evidence in high-resolution images, Visual Sketchpad incorporates visual tool calls and their outputs, and ICoT organizes reasoning through interleaved image patches and text (Wu & Xie, 2024; Hu et al., 2024; Gao et al., 2025). DeepEyes and Pixel Reasoner learn visual tool use through reinforcement learning, while Mini-o3 expands the reasoning patterns and interaction turns used in visual search (Zheng et al., 2026; Wang et al., 2025a; Lai et al., 2026). These methods advance decisions about when and where to look. We focus on the computation after a crop request: retaining the existing observation policy, reusing previously computed regional states, and adapting their use to the current reasoning objective.

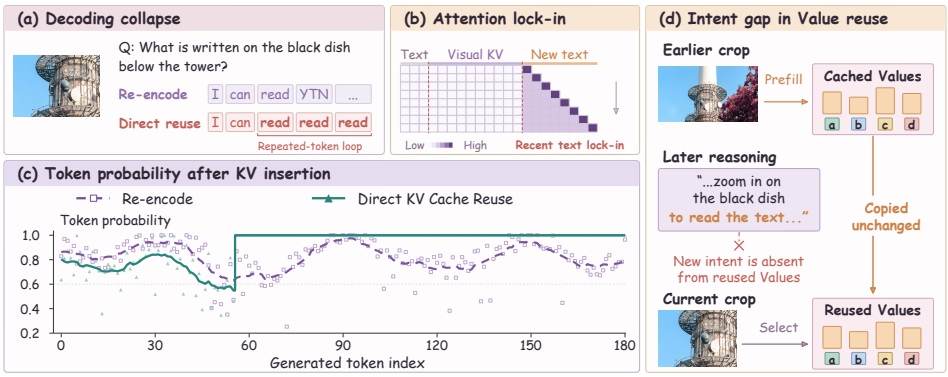


Figure 2: **Why visual KV reuse requires adaptation.** (a,b) Outputs and attention patterns illustrate repetitive decoding and attention concentrated on recent text under direct reuse. (c) Generated-token probabilities abruptly approach one under direct reuse but remain variable after full re-encoding (Re-encode). (d) Skipping visual prefill leaves reused values without the intervening crop intent.

2.2 Efficient Inference and KV Cache Reuse

FastV, SparseVLM, VisionZip, and LLaVA-PruMerge reduce visual-token computation through pruning or aggregation (Chen et al., 2024; Zhang et al., 2025b; Yang et al., 2025; Shang et al., 2025), while VASparse combines visual sparsification with hallucination mitigation (Zhuang et al., 2025). H_2O, StreamingLLM, and SnapKV retain cache states based on attention, attention sinks, and prompt relevance, respectively (Zhang et al., 2023; Xiao et al., 2024; Li et al., 2024b). For changing contexts, CacheBlend selectively recomputes cached text chunks (Yao et al., 2024); MPIC combines position-independent multimodal caching with selective recomputation (Zhao et al., 2025); Kamera adds low-rank contextual compensation to positional relocation (Ma et al., 2026); and CONDUIT uses attention and values to select visual states for refresh (Chen et al., 2026). We focus on regional visual KV reuse across successive crops, adapting cached states through adaptive routing, key-position rebinding, and intent-guided value enhancement while retaining full re-encoding when new detail is needed.

3 Observations and Motivation

Successive crops often cover overlapping image regions, creating an opportunity to reuse visual KV states and avoid repeated computation. However, these states were generated in an earlier reasoning context, and whether they can be directly reused for a later crop remains unclear. We therefore begin by examining the behavior of direct reuse and skipped visual prefill in Figure 2 to identify the conditions needed for effective visual cache reuse.

3.1 Stale Positional Bindings and Decoding Collapse

Direct insertion of historical visual KV can lead to repetitive token loops (Figure 2(a,b)). To understand this behavior, we examine how the position encoded in a cached key affects attention scores. Omitting layer and head indices, let $\mathbf{p}_j^{\mathrm{src}}$ be the source position of selected visual token j and $\bar{\mathbf{k}}_j$ its key before rotation. The historical cache stores $\mathbf{k}_j^{\mathrm{src}} = \mathbf{R}(\mathbf{p}_j^{\mathrm{src}})\bar{\mathbf{k}}_j$. For a subsequent text token at position $\mathbf{p}_u$, with pre-rotation query $\bar{\mathbf{q}}_u$, direct reuse gives

$$s_{u,j}^{\mathrm{direct}} = \frac{\big(\mathbf{R}(\mathbf{p}_u)\bar{\mathbf{q}}_u\big)^\top \mathbf{R}(\mathbf{p}_j^{\mathrm{src}})\bar{\mathbf{k}}_j}{\sqrt{d}} = \frac{\bar{\mathbf{q}}_u^\top \mathbf{R}(\mathbf{p}_j^{\mathrm{src}} - \mathbf{p}_u)\bar{\mathbf{k}}_j}{\sqrt{d}}. \tag{1}$$

Here d is the head dimension and $\mathbf{R}(\cdot)$ is the RoPE rotation operator. Equation (1) shows that relocating a cache entry does not change the position encoded within its key. At the current insertion position $\mathbf{p}_j^{\mathrm{new}}$, the query–key interaction requires the relative rotation $\mathbf{R}(\mathbf{p}_j^{\mathrm{new}} - \mathbf{p}_u)$, whereas direct reuse still uses $\mathbf{R}(\mathbf{p}_j^{\mathrm{src}} - \mathbf{p}_u)$. This mismatch can alter attention matching even when the content projections remain fixed.

The resulting shift also affects the readout from the entire context. Let $\mathcal{C}_u$ denote the full context visible to the current token, including historical text and the newly inserted visual KV states. Attention weights and their output are (with $s_{u,i} = \mathbf{q}_u^\top \mathbf{k}_i / \sqrt{d}$, where $\mathbf{q}_u$ and $\mathbf{k}_i$ include RoPE, $\mathbf{v}_i$ is the value at context index i, and $\mathbf{o}_u$ is the head output)

$$\alpha_{u,i} = \frac{\exp(s_{u,i})}{\sum_{v \in \mathcal{C}_u} \exp(s_{u,v})}, \qquad \mathbf{o}_u = \sum_{i \in \mathcal{C}_u} \alpha_{u,i} \mathbf{v}_i. \tag{2}$$

All attention weights in Equation (2) share a softmax denominator, so changing reused visual-key scores also redistributes attention to reasoning text. This affects next-token predictions and may accumulate as generated tokens enter the context, reinforcing recent outputs. Figure 2(a,b) shows token repetition and attention lock-in; Figure 2(c) contrasts probabilities approaching one under direct reuse with continued variation after full re-encoding. These effects motivate correcting positions while retaining cached key content. Section 4.1 introduces Positional Key Rebinding (PKR), which uses RoPE's rotation properties to encode clean cached keys at their current insertion positions.

3.2 The Semantic Gap Left by Skipping Visual Prefill

Position correction aligns old keys with new positions but does not restore prefill's semantic interaction. After inspecting a tower, for example, the model may request a crop of the black dish below it to read its text (Figure 2(d)). Full prefill places the crop's visual tokens after this reasoning, allowing the new objective to influence later-layer representations through causal attention. Direct value reuse skips this interaction.

Let $(\mathbf{E}_{\text{src}}, \mathcal{P}_{\text{src}})$ denote the source observation's visual input and preceding context, and $(\mathbf{E}_{\text{crop}}, \mathcal{P}_{\text{crop}})$ those of the current crop. Only $\mathcal{P}_{\text{crop}}$ includes the intervening reasoning. Let $\psi(j)$ map crop token j to its source-cache index. Omitting layer and head indices, the value projection $\mathbf{W}_V$ acts on pre-projection visual states $\mathbf{h}$, giving

$$\begin{aligned} \mathbf{v}_j^{\text{prefill}} &= \mathbf{W}_V \mathbf{h}_j(\mathbf{E}_{\text{crop}}; \mathcal{P}_{\text{crop}}), \\ \mathbf{v}_j^{\text{reuse}} &= \mathbf{v}_j^0 = \mathbf{W}_V \mathbf{h}_{\psi(j)}(\mathbf{E}_{\text{src}}; \mathcal{P}_{\text{src}}). \end{aligned} \tag{3}$$

Here $\mathbf{v}_j^{\text{prefill}}$ is newly computed for crop token j, $\mathbf{v}_j^{\text{reuse}}$ is its directly reused value, and $\mathbf{v}_j^0$ is the unchanged value retrieved from the source cache. These cached values predate the new crop intent and may originate from an older fully prefilled ancestor. Current queries can adjust attention weights but still read these earlier representations. Section 4.2 introduces Intent-guided Value Enhancement (IVE): pre-crop reasoning attention strengthens cached values relevant to the current objective without visual prefill.

4 Method

Our method handles each crop request by deciding whether new visual information is needed from the original image, then supplying the corresponding visual KV states for subsequent reasoning. As shown in Figure 3(a), the framework has two paths. **Re-encoding** crops the target region from the original image and performs visual encoding and prefill. **Cache reuse** retrieves existing visual states for the region, adapts them through Positional Key Rebinding (PKR) and Intent-guided Value Enhancement (IVE), and resumes text generation.

Adaptive routing. At crop step t, the model requests region $B_t \subseteq \Omega_s$ from observation O_s ($s < t$), which we call the parent observation. Routing depends on the crop's area relative to its parent (Figure 3(a)). Smaller crops typically require further magnification, so we crop and re-encode original-image pixels to obtain finer visual detail. Larger crops preferentially reuse available states to reduce repeated processing of overlapping content. Let Ω_s denote the parent's valid image region and $H_t \in \{0, 1\}$ indicate whether the required visual cache entries and regional mappings are available. Both areas are measured in the same parent-image coordinates, with $|\Omega_s| > 0$. We select the path as

$$r_t = \frac{|B_t|}{|\Omega_s|}, \qquad z_t = \mathbf{1}[r_t > \tau \ \wedge \ H_t = 1]\,. \tag{4}$$

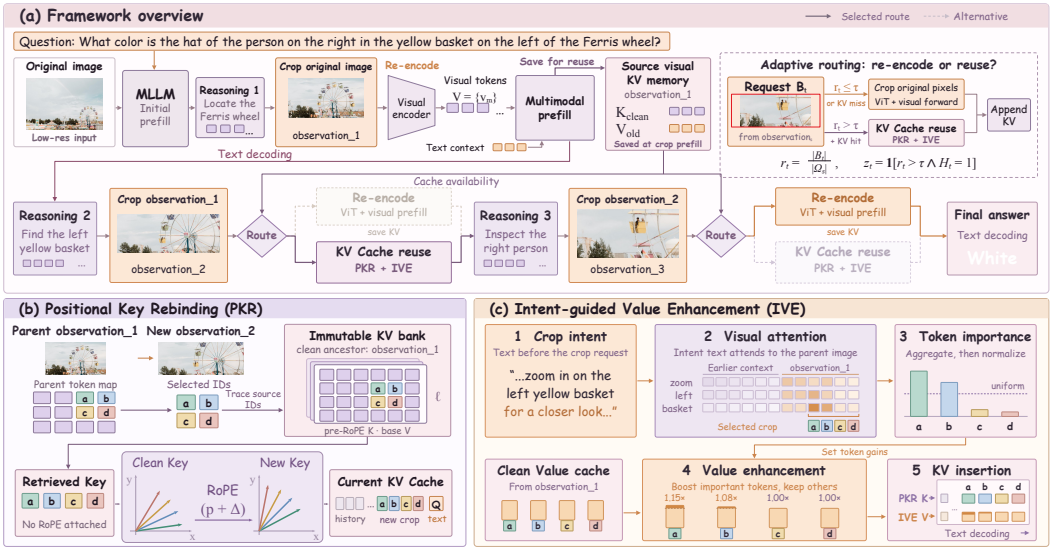


Figure 3: **Adaptive visual KV cache reuse.** (a) Requests are routed to full re-encoding or regional cache reuse. Fully re-encoded crops populate an immutable source bank. (b) PKR traces selected tokens to clean source keys and applies RoPE at the target positions $\mathbf{p}_j^{\text{new}}$. (c) IVE uses pre-crop reasoning attention to derive token-wise gains and enhance base values before insertion. Subsequent text decoding proceeds normally.

Here $|\cdot|$ denotes area, $\mathbf{1}[\cdot]$ is the indicator function, and τ is the area-ratio threshold; $z_t = 1$ selects cache reuse. When $r_t \leq \tau$ or the required cache is unavailable, we use full re-encoding: map the target region to the original image, crop and magnify its pixels, and perform visual encoding and visual-token prefill.

Saving and retrieving visual KV caches. As shown in Figure 3(a,b), full re-encoding stores clean pre-RoPE keys, unenhanced base values, and regional mappings at every layer. For each crop, we trace these mappings to its fully prefilled source, including an ancestor if the parent was itself obtained through reuse. Let $j = 1, \ldots, N$ index the selected visual tokens. We write $\bar{\mathbf{k}}_j$ and $\mathbf{v}_j^0$ for the retrieved entries, already aligned to crop token j through $\psi(j)$. The source bank remains unchanged; each reuse adapts a fresh retrieval of these base states.

For clarity, the following token-level equations fix one crop step, decoder layer, and KV head, and omit these indices. The operations are applied independently at every layer and KV head; query-head indices are retained where needed. Appendix A.3 summarizes the symbols.

4.1 Positional Key Rebinding (PKR)

Directly reused keys retain old positional encodings, creating a mismatch with subsequent text queries (Section 3.1). PKR corrects this mismatch by shifting selected visual tokens to the current insertion positions through a shared RoPE offset.

Let $\mathbf{p}_j^{\text{src}}$ be the source RoPE position of selected token j, and $\mathbf{p}_{\text{ins}}$ the starting position of the new visual KV block. Anchoring the first selected token there gives the common offset $\boldsymbol{\Delta} = \mathbf{p}_{\text{ins}} - \mathbf{p}_1^{\text{src}}$ and target positions $\mathbf{p}_j^{\text{new}} = \mathbf{p}_j^{\text{src}} + \boldsymbol{\Delta}$. We apply RoPE to each retrieved clean key at its new position (Figure 3(b)):

$$\mathbf{k}_j^{\text{new}} = \mathbf{R}(\mathbf{p}_j^{\text{new}})\bar{\mathbf{k}}_j, \tag{5}$$

where $\mathbf{k}_j^{\text{new}}$ is the rebound key and $\mathbf{R}(\cdot)$ is the model's rotary position operator. For MRoPE, positions and offsets use the model's native multidimensional coordinates (Bai et al., 2025).

The common shift preserves relative positions among the selected visual tokens. This invariance assumes a shared offset in the model's RoPE coordinates; it does not hold in general for arbitrary token-wise reindexing. By RoPE's rotation invariance, the shifted keys retain the norms and pairwise inner products of the keys encoded at their source positions (Su et al., 2021), while updating their

positional relation to current text queries. PKR thus corrects the mismatch in Section 3.1 while preserving the computed key content. Each reuse starts from clean keys and applies RoPE once at $\mathbf{p}_j^{\text{new}}$, without recomputing content projections.

4.2 Intent-guided Value Enhancement (IVE)

After correcting positions, we also account for why the crop is requested. Reasoning between observations typically identifies the next object, region, or attribute to inspect. Full prefill lets new visual tokens interact with this reasoning, whereas reuse retrieves values formed during an earlier observation (Section 3.2). IVE uses the pre-crop reasoning's attention to the requested region B_t within its immediate parent observation O_s to form a regional relevance prior and enhance the corresponding cached values (Figure 3(c)).

Reading visual attention from pre-crop reasoning. Let $\mathcal{I}$ be the reasoning-token window preceding the current crop request, and $\mathcal{H}$ the query heads sharing the fixed KV head. For $u \in \mathcal{I}$ and $h \in \mathcal{H}$, query $\mathbf{q}_{h,u}$ attends to selected visual token j within B_t at parent-context index x_j. For $x_j \in \mathcal{C}_u$, the attention weight is

$$a_{h,u,j} = \frac{\exp\left(\mathbf{q}_{h,u}^\top \mathbf{k}_{x_j}/\sqrt{d}\right)}{\sum_{i \in \mathcal{C}_u} \exp\left(\mathbf{q}_{h,u}^\top \mathbf{k}_i/\sqrt{d}\right)}. \tag{6}$$

Here $a_{h,u,j}$ is the attention weight saved during decoding, d is the query/key head dimension, and $\mathcal{C}_u$ is the full causal context visible to u. We set $a_{h,u,j} = 0$ when $x_j \notin \mathcal{C}_u$. The queries and keys already include RoPE. The index x_j locates the token in O_s's visual block in the active context; it is distinct from the source-cache index $\psi(j)$ and the token's RoPE coordinates. Full-context normalization retains competition with other visual content and reasoning text.

Forming the regional relevance prior. For a nonempty reasoning window with positive attention mass on B_t, we average over reasoning tokens and query heads sharing the fixed KV head, then normalize across the $N \geq 1$ selected visual tokens within B_t:

$$b_j = \frac{\sum_{u \in \mathcal{I}} \sum_{h \in \mathcal{H}} a_{h,u,j}}{|\mathcal{I}|\,|\mathcal{H}|}, \qquad \pi_j = \frac{b_j}{\sum_{i=1}^{N} b_i}. \tag{7}$$

Here b_j is the aggregated attention and π_j the token's relevance within B_t. If the reasoning window is empty or regional attention mass is zero or unavailable, we use $\pi_j = 1/N$, yielding no enhancement. Aggregating before regional normalization gives greater influence to reasoning text that attends strongly to the target region, avoiding equal weight for text that barely attends to it.

Enhancing values relevant to the current target. Using the uniform distribution $1/N$ as a baseline, we convert above-baseline relevance into a gain G_j and apply it to the retrieved base value:

$$G_j = 1 + \beta\ \mathrm{clip}(N\pi_j - 1,\, 0,\, 1)\,, \qquad \mathbf{v}_j^{\text{new}} = G_j \mathbf{v}_j^0. \tag{8}$$

Here $\mathrm{clip}(x, 0, 1) = \min(1, \max(0, x))$, and $\beta \geq 0$ controls enhancement, with $1 \leq G_j \leq 1 + \beta$. Only tokens above the uniform relevance baseline are enhanced; the others remain unchanged. Every reuse applies its gain to the unchanged base value $\mathbf{v}_j^0$, avoiding cumulative amplification. Finally, the keys $\mathbf{k}_j^{\text{new}}$ and values $\mathbf{v}_j^{\text{new}}$ from Equations (5) and (8) are inserted into the current cache, and decoding resumes.

Table 1: **Results on VisualProbe and HR-Bench 8K.** Accuracy (%; higher is better). Shaded rows compare Mini-o3, its ablation without intermediate crop images (*w/o Interleaved images*), and our cache-reuse method. All shaded results report Avg@8, averaged over eight runs. Bold and underlining mark column-wise best and second-best results (including ties).

Model	VisualProbe				HR-Bench 8K		
	Hard	Medium	Easy	**Overall**	**FSP**	**FCP**	**Overall**
LLaVA-OneVision Li et al. (2024a)	13.4	12.5	36.2	19.2	67.3	52.3	59.8
Qwen2.5-VL Bai et al. (2025)	23.9	26.0	39.1	29.2	78.8	51.8	65.3
Pixel-Reasoner Wang et al. (2025a)	28.8	29.6	58.4	37.3	80.0	54.3	66.9
Thyme Zhang et al. (2026)	46.2	43.3	62.4	49.1	86.5	57.5	72.0
DeepEyes Zheng et al. (2026)	35.1	29.8	60.1	39.2	86.8	**58.5**	72.6
Mini-o3 Lai et al. (2026)	49.1	**49.6**	68.1	54.6	87.5	56.1	71.8
w/o Interleaved images	*18.9*	*23.5*	*50.4*	*29.9*	*70.3*	*55.0*	*62.6*
Mini-o3 + PKR + IVE (Ours)	**50.0**	48.5	**70.2**	**54.8**	**87.7**	**58.5**	**73.1**

Table 2: **Results on MME-RealWorld-Lite, V* Bench, and HR-Bench 4K.** Accuracy (%; higher is better). Shaded rows compare Mini-o3, its ablation without intermediate crop images (*w/o Interleaved images*), and our cache-reuse method. All shaded results report Avg@8, averaged over eight runs. Bold and underlining mark column-wise best and second-best results (including ties).

Model	MME-RealWorld-Lite			V^* Bench			HR-Bench 4K		
	Perception	Reasoning	**Overall**	**Attr**	**Spatial**	**Overall**	**FSP**	**FCP**	**Overall**
LLaVA-OneVision Li et al. (2024a)	52.8	42.7	48.5	75.7	75.0	75.4	72.0	54.0	63.0
Qwen2.5-VL Bai et al. (2025)	48.8	37.7	44.1	73.9	67.1	71.2	85.2	52.2	68.8
Pixel-Reasoner Wang et al. (2025a)	53.1	44.5	49.7	83.5	76.3	80.6	86.0	60.3	72.9
Thyme Zhang et al. (2026)	59.1	49.1	55.2	83.5	80.3	82.2	91.0	63.0	77.0
DeepEyes Zheng et al. (2026)	56.6	50.5	54.3	**91.3**	**88.2**	**90.1**	91.3	59.0	75.1
Mini-o3 Lai et al. (2026)	66.8	59.9	64.1	89.4	82.9	86.8	89.5	**63.7**	76.6
w/o Interleaved images	*53.4*	*46.5*	*50.7*	*82.6*	*76.3*	*80.1*	*81.5*	*63.3*	*72.4*
Mini-o3 + PKR + IVE (Ours)	**67.2**	**60.3**	**64.5**	87.8	86.8	87.4	**91.7**	62.7	**77.2**

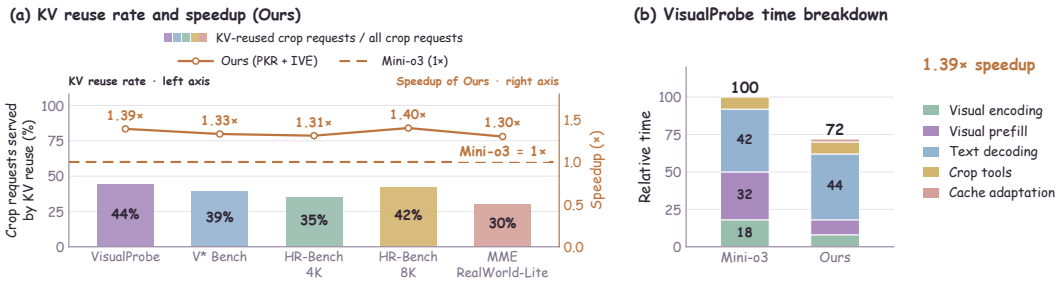


Figure 4: **Inference efficiency of our method.** (a) KV reuse rate (% of all crop requests) and end-to-end speedup of Ours over Mini-o3. (b) VisualProbe time breakdown, normalized to Mini-o3 = 100.

5 Experiments

5.1 Experimental Setup

Benchmarks. We use VisualProbe (Lai et al., 2026) as our primary benchmark for complex visual search, reporting accuracy on the Hard, Medium, and Easy splits as well as overall accuracy. We also evaluate high-resolution visual understanding on V* Bench (Wu & Xie, 2024) and HR-Bench 4K/8K (Wang et al., 2025b), and general perception and reasoning on MME-RealWorld-Lite (Zhang et al., 2025a).

Baselines. We use Mini-o3 (Lai et al., 2026) as the sole baseline for direct comparison. Prior work reports nearly unchanged accuracy for DeepEyes, Thyme, and Pixel Reasoner when tool-generated image outputs are suppressed while retaining the preceding textual reasoning (Yang et al., 2026). In such settings, preserving accuracy while accelerating inference provides limited evidence that KV reuse retains useful intermediate visual information. Our ablation shows that removing cropped images causes a substantial performance drop for Mini-o3. We therefore implement our method on Mini-o3 to evaluate whether cache reuse reduces repeated re-encoding while preserving multi-turn reasoning that depends on intermediate visual information.

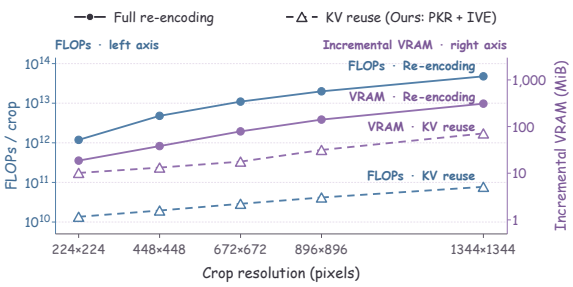


Figure 5: **Crop processing costs on Mini-o3.** Full re-encoding versus PKR + IVE: FLOPs (left) and incremental VRAM (MiB, right), both on log scales.

Table 3: **Incremental ablation** (accuracy, %). Base: Mini-o3 without intermediate crop images; subsequent rows add PKR-based reuse and IVE. Bold marks column-wise best results.

Base	PKR	IVE	VisualProbe Overall	V* Bench Overall	HR-Bench 8K Overall
✓			29.9	80.1	62.6
✓	✓		50.4	83.6	71.9
✓	✓	✓	**54.8**	**87.4**	**73.1**

Implementation Details. We implement our method in Mini-o3's multi-turn vLLM (Kwon et al., 2023) pipeline with frozen weights and no additional training. Routing uses the crop-to-parent area ratio and cache availability, with $\tau = 0.2$. We set $\beta = 0.15$ for IVE, bounding value gains within $[1, 1.15]$. Additional implementation and evaluation settings are provided in Appendix A.2.

5.2 Main Results

Tables 1 and 2 show that our method preserves or improves Mini-o3's overall accuracy across all five benchmarks, with gains of 0.2–1.3 percentage points. Removing intermediate crop images substantially degrades performance, including a 24.7-point drop on VisualProbe, confirming the importance of retaining intermediate visual evidence.

Figure 4(a) shows that serving 30–44% of crop requests through KV reuse yields 1.30–$1.40\times$ end-to-end speedups. Because crop requests differ in processing cost, the reuse ratio alone does not determine the end-to-end speedup. On VisualProbe, visual encoding and prefill together decrease from 50 to 18 normalized time units, while cache adaptation adds only 2 units. These savings outweigh the slightly longer text decoding, reducing total inference time by 28% (Figure 4(b)). Figure 5 further shows that PKR + IVE reduces per-crop FLOPs and incremental VRAM relative to full re-encoding across all tested resolutions. Both paths become more costly as crop resolution increases, but the FLOPs gap widens in favor of reuse. Incremental VRAM also remains lower for reuse across the tested range, showing that the compute savings do not require higher per-crop memory use.

5.3 Ablation Studies

To analyze the contributions of PKR and IVE, we start from Mini-o3 with intermediate crop images removed and progressively add the two components, as shown in Table 3. Enabling PKR-based visual KV reuse raises accuracy from 29.9% to 50.4% on VisualProbe, 80.1% to 83.6% on V* Bench, and 62.6% to 71.9% on HR-Bench 8K, recovering much of the performance lost without intermediate images. With routing unchanged, IVE adds 4.4, 3.8, and 1.2 percentage points, respectively, showing that current crop intent helps the model better use cached visual evidence.

5.4 Qualitative Analysis

Successive cropping with cache reuse. Figure 6 illustrates two examples of text and object recognition. In both cases, the first crop establishes the visual cache, which supports the next two crop requests through PKR and IVE without repeating visual encoding or visual-token prefill. Although the third crop is requested from the second observation, its visual states are retrieved from the cache established by the first crop. The fourth crop returns to the original pixels for full re-encoding to obtain finer detail. The resulting answers, "YTN" and "air conditioner," match the ground truth, illustrating how reuse supports intermediate inspection while re-encoding provides additional detail for the final decision.

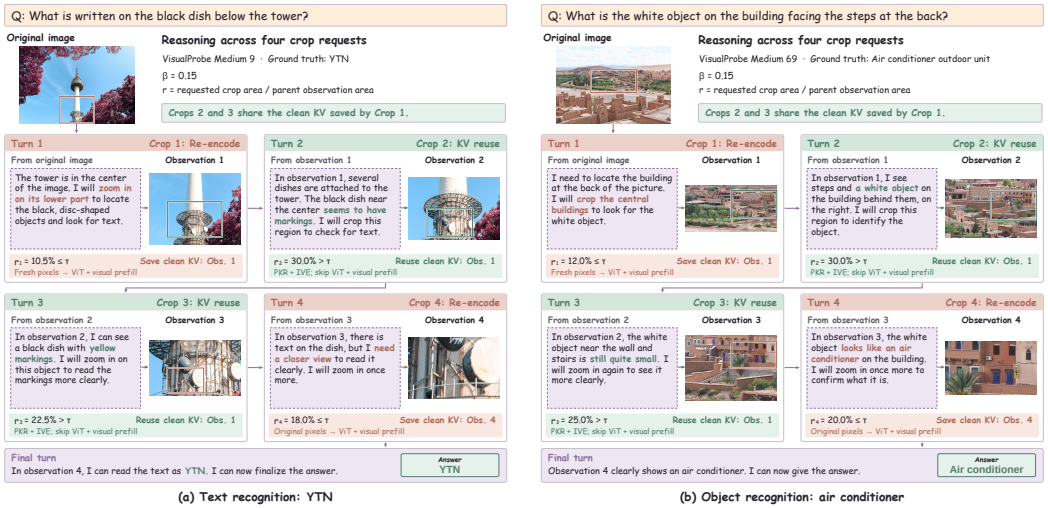


Figure 6: **Successive cropping with cache reuse.** (a) Reading text on a dish. (b) Identifying an air conditioner. Both examples use four crop requests, with KV reuse at the second and third steps.

Intent-guided value enhancement. Figure 7 illustrates a multi-turn search for the letters on a relief beneath a flag, where successive crops narrow the view from the building to the target region. Our framework re-encodes Crops 1 and 3 to obtain finer visual details and reuses cached KV states for Crops 2 and 4 to avoid repeated computation. The fourth crop further highlights the role of IVE: although the cached visual evidence already contains the letters USA, the model with PKR alone still requests another crop; with IVE, it instead answers USA using the available evidence. The visualizations show that the pre-crop reasoning intent to zoom in and read the letters produces a relevance prior that emphasizes the letter-bearing region. IVE converts this prior into bounded gains on the corresponding cached values, strengthening their contributions during subsequent decoding. This example illustrates that the presence of relevant visual information in the cache does not guarantee its effective use: adapting reused values to the current reasoning intent can help the model extract an answer from existing evidence.

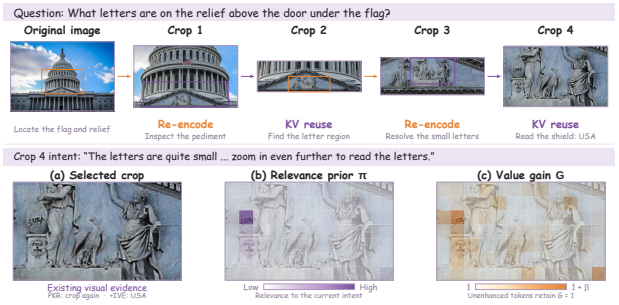


Figure 7: **Adaptive routing and intent-guided value enhancement.** Top: successive crops alternate between re-encoding and KV reuse. Bottom: (a) the selected fourth crop, (b) the intent-derived relevance prior π, and (c) the resulting value gains G. The model with PKR alone requests another crop; adding IVE yields the answer USA.

6 Conclusion

We present an adaptive visual KV cache reuse framework for efficient multi-turn multimodal reasoning. The framework selects between regional cache reuse and full re-encoding, reducing repeated visual processing while preserving access to additional image detail. Our analysis connects two challenges arising from skipped visual prefill: stale positional bindings in cached keys and the absence of intervening crop intent in historical values. Positional Key Rebinding corrects the former by assigning cached keys their current positions, while Intent-guided Value Enhancement addresses the latter by modulating value contributions using pre-crop reasoning attention. These adaptations require no additional training and leave the model's crop-decision mechanism unchanged. On reuse hits, the framework resumes reasoning from adapted visual states without repeating visual encoding or decoder prefill for the selected visual tokens.

AI USE STATEMENT

We used generative AI tools to assist with Chinese-to-English translation, manuscript drafting and revision, refining the exposition of the method and its mathematical analysis, literature search, reference formatting, and figure layout refinement. The authors reviewed the AI-assisted content and verified the technical statements and references. The authors take full responsibility for the final manuscript, including all claims, results, and supporting materials.

REFERENCES

Shuai Bai et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. doi: 10.48550/arXiv.2502.13923.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, 2024.

Pengan Chen, Kaisheng Zheng, Liang Hong, Lixia Yi, Jiyue Jiang, Jiayang Chen, Yixuan Wang, Yimin Fan, Xinyuan Liu, Jiayi Li, Zhanqiu Zhang, Yiwen Guo, and Yu Li. CONDUIT: A unified residual-stream restoration framework for KV cache reuse in vision-language models. *arXiv preprint arXiv:2609.05821*, 2026. doi: 10.48550/arXiv.2609.05821.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19520–19529, 2025.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. Visual Sketchpad: Sketching as a visual chain of thought for multimodal language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165.

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*, 2026.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a. doi: 10.48550/arXiv.2408.03326.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024b.

Bole Ma, Jan Eitzinger, Harald Koestler, and Gerhard Wellein. Kamera: Unified position-invariant multimodal KV cache for training-free reuse. *arXiv preprint arXiv:2606.23581*, 2026. doi: 10.48550/arXiv.2606.23581.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22857–22867, 2025.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. doi: 10.48550/arXiv.2104.09864.

Haozhe Wang, Alex Su, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel-space reasoning with curiosity-driven reinforcement learning. *arXiv preprint arXiv:2505.15966*, 2025a.

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, 2025b.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. VisionZip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19792–19802, 2025.

Wenjie Yang, Siqi Zhu, and Zengfeng Huang. Position: Your VLM may not be thinking with interleaved images. In *Forty-third International Conference on Machine Learning Position Paper Track*, 2026.

Jiayi Yao, Hanchen Li, Yuhan Liu, Siddhant Ray, Yihua Cheng, Qizheng Zhang, Kuntai Du, Shan Lu, and Junchen Jiang. CacheBlend: Fast large language model serving for RAG with cached knowledge fusion. *arXiv preprint arXiv:2405.16444*, 2024. doi: 10.48550/arXiv.2405.16444.

Peng Ye, Tong He, Shengji Tang, Baopu Li, Tao Chen, Lei Bai, and Wanli Ouyang. Stimulative Training++: Go beyond the performance limits of residual networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(8):6463–6479, August 2025a. doi: 10.1109/TPAMI.2025.3558421.

Peng Ye, Chenyu Huang, Mingzhu Shen, Tao Chen, Yongqi Huang, and Wanli Ouyang. Dynamic model merging with mixture of weights. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(8):7925–7939, August 2025b. doi: 10.1109/TCSVT.2025.3544680.

Peng Ye, Tong He, Baopu Li, Tao Chen, Lei Bai, and Wanli Ouyang. β-DARTS++: Bi-level regularization for proxy-robust differentiable architecture search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1407–1423, February 2026. doi: 10.1109/TPAMI.2025.3616249.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. MME-RealWorld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations*, 2025a.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, Haonan Fan, Kaibing Chen, Jiankang Chen, Haojie Ding, Kaiyu Tang, Zhang Zhang, Liang Wang, Fan Yang, Tingting Gao, and Guorui Zhou. Thyme: Think beyond images. In *International Conference on Learning Representations*, 2026.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A. Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. SparseVLM: Visual token sparsification for efficient vision-language model inference. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 74840–74857, 2025b.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H_2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1506.

Shiju Zhao, Junhao Hu, Rongxiao Huang, Jiaqi Zheng, and Guihai Chen. MPIC: Position-independent multimodal context caching system for efficient MLLM serving. *arXiv preprint arXiv:2502.01960*, 2025. doi: 10.48550/arXiv.2502.01960.

Ziwei Zheng, Minghao Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. DeepEyes: Incentivizing "thinking with images" via reinforcement learning. In *International Conference on Learning Representations*, 2026.

Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, and Yuexian Zou. VASparse: Towards efficient visual hallucination mitigation via visual-aware token sparsification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4189–4199, 2025.

A Appendix

A.1 Supplementary Material Overview

We organize the supplementary material as follows:

- Additional Implementation (Section A.2).
- Notation Summary (Section A.3).
- Algorithm Summary (Section A.4).
- Visualization of Intent-Guided Value Enhancement (Section A.5).
- Visualization of Multi-Turn KV Cache Reuse (Section A.6).
- Limitations and Future Work (Section A.7).

A.2 Additional Implementation

Test Benchmarks. We evaluate on five benchmark settings covering iterative visual search, fine-grained high-resolution perception, and general visual reasoning. These include VisualProbe, V* Bench, the 4K and 8K subsets of HR-Bench, and MME-RealWorld-Lite, as described below.

VisualProbe (Lai et al., 2026) evaluates challenging visual search in high-resolution images containing small targets and numerous distractors. Finding the evidence needed to answer a question can require successive local inspections, exploratory cropping, and revision of earlier search decisions. We use it as our primary benchmark for assessing whether cached visual states can support multi-turn reasoning as the inspection objective changes. We report accuracy on the Hard, Medium, and Easy splits, together with Overall accuracy.

V* Bench (Wu & Xie, 2024) focuses on fine-grained visual understanding in 191 high-resolution images. Its questions require recognizing object attributes, such as color and material, or determining spatial relationships between objects. Answering them depends on locating the relevant objects and inspecting their details within visually complex scenes. This benchmark tests whether reuse preserves the visual evidence needed for attribute recognition and spatial reasoning. We report Attr, Spatial, and Overall accuracy.

HR-Bench 4K (Wang et al., 2025b) evaluates perception using 4K images cropped around relevant objects from the original 8K scenes. It covers Fine-grained Single-instance Perception (FSP), including attribute recognition and text reading, and Fine-grained Cross-instance Perception (FCP), including spatial relationships and reasoning over maps or charts. We use this setting to assess whether regional cache reuse retains local detail and the relationships needed for answering. Results are reported as FSP, FCP, and Overall accuracy.

HR-Bench 8K (Wang et al., 2025b) uses the original 8K images and the same FSP and FCP task categories. Compared with the cropped 4K setting, these images retain a larger field of view, requiring relevant details to be located within a broader visual context. This setting evaluates adaptive reuse when the model must balance revisiting existing evidence with obtaining finer detail through re-encoding. We report FSP, FCP, and Overall accuracy separately from the 4K results.

MME-RealWorld-Lite (Zhang et al., 2025a) is a compact subset of MME-RealWorld for evaluating perception and reasoning in practical visual scenarios. It covers optical character recognition in the wild, remote sensing, diagrams and tables, monitoring, and autonomous driving. These tasks involve reading visual content, identifying objects, and reasoning about information in complex images. It complements the targeted visual search benchmarks by testing our method across a broader range of visual tasks. We report Perception, Reasoning, and Overall accuracy.

Inference and cache reuse. We use the frozen Mini-o3-7B-v1 model and handle model-issued crop requests within a multi-turn vLLM inference pipeline. We reuse visual KV states across all 28 layers when the crop-to-parent area ratio exceeds the predefined threshold and the required cache entries are available; otherwise, we fully re-encode the crop. During reuse, we rebind cached keys to the visual positions of the new crop and apply the current gains to the original cached values, preventing enhancement from accumulating across consecutive reuse steps.

Intent-guided value enhancement. IVE uses up to the last 64 generated tokens preceding the crop request to represent the current reasoning intent. We aggregate their attention to the requested crop region within its immediate parent observation and normalize within that region to obtain the relevance prior for bounded value gains. We set $\beta = 0.15$ and use FP32 for IVE attention computation and normalization. Model parameters remain frozen. After processing the crop, the model resumes ordinary text generation.

Evaluation settings. We sample one answer per question in each evaluation run and repeat the full evaluation eight times. Avg@8, reported for the shaded comparisons in Tables 1 and 2, is the arithmetic mean of the accuracies from these eight runs. A local Qwen2.5-72B Base model judges each final answer given the question, reference answer, and model's final response, and returns a binary correctness score. Table 4 summarizes the configuration.

Table 4: **Implementation and evaluation configuration.**

Item	Configuration
Inference environment	vLLM 0.9.2; 8 NVIDIA RTX 5880 Ada GPUs, each with 48 GiB
Numerical precision	BF16 model inference; FP32 IVE attention computation and normalization
Decoding	Temperature 1.0, top-p 1.0, top-k -1; at most 12 rounds
Image and intent settings	Image limit: 2 million pixels; up to 64 pre-crop tokens for intent; $\beta = 0.15$
Answer judge	Qwen2.5-72B Base; BF16; 8-way tensor parallelism; 64 questions per batch

A.3 NOTATION SUMMARY

Table 5 summarizes the notation used in the main text and Algorithm 1. Unless explicitly shown, crop-step, layer, and KV-head indices are omitted. PKR and IVE operate independently at every layer and KV head; IVE averages only over the intent tokens and query heads sharing that KV head. Source states remain unchanged, and each reuse applies its gain to the base value.

Table 5: **Main notation.** Subscripts src and crop denote the fully prefilled source and current crop.

Symbol	Definition
Routing	
I, Q	Original image and input question.
$O_s, O_t;\ s < t$	Immediate parent observation and current crop; their observation indices.
B_t, Ω_s	Requested crop region and the parent's valid image region, in the same coordinates.
r_t	Crop-to-parent area ratio, $\lvert B_t\rvert/\lvert\Omega_s\rvert$.
H_t, z_t	Cache-availability indicator and routing decision; $z_t = 1$ selects reuse.
τ	Area-ratio threshold; reuse requires $r_t > \tau$ and $H_t = 1$.
Cache and position	
j, N	Index and number of selected visual tokens within B_t in O_s, $j = 1, \ldots, N$.
$\psi(j), x_j$	Token j's source-cache index and its sequence index in the parent context, respectively.
$\mathbf{E}_{\mathrm{src}}, \mathbf{E}_{\mathrm{crop}}$, $\mathcal{P}_{\mathrm{src}}, \mathcal{P}_{\mathrm{crop}}$	Source/crop visual inputs and preceding contexts. The source may be an earlier ancestor of the parent.
$\mathbf{W}_V, \mathbf{h}_j(\mathbf{E}; \mathcal{P})$	Value-projection matrix and token j's state entering that projection at the fixed layer.
$\mathbf{v}_j^{\mathrm{prefill}}$	Value computed by full prefill of the current crop under its current context.
$\mathbf{v}_j^{\mathrm{reuse}}$	Directly reused value, equal to $\mathbf{v}_j^0$, before IVE adaptation.
$\mathbf{v}_j^0$	Unenhanced base value retrieved through $\psi(j)$; never overwritten by reuse gains.
$\bar{\mathbf{k}}_j, \bar{\mathbf{q}}_u$	Retrieved clean source key and text query before RoPE.
$\mathbf{k}_j^{\mathrm{src}}, \mathbf{k}_j^{\mathrm{new}}$	Selected key encoded at its source position and rebound to its new position.
$\mathbf{p}_j^{\mathrm{src}}, \mathbf{p}_j^{\mathrm{new}}, \mathbf{p}_u$	Source and target visual-token coordinates, and text-token RoPE position; distinct from sequence indices.
$\mathbf{p}_{\mathrm{ins}}, \boldsymbol{\Delta}$	Insertion anchor and shared positional offset, $\boldsymbol{\Delta} = \mathbf{p}_{\mathrm{ins}} - \mathbf{p}_1^{\mathrm{src}}$.
$\mathbf{R}(\cdot), d$	RoPE rotation operator and query/key head dimension.
Intent and attention	
$u, i, \mathcal{C}_u$	Query-token index, context-token index, and full causal context visible to u.
$\mathbf{q}_u, \mathbf{k}_i, \mathbf{v}_i$	Query, context key, and context value; queries and keys include RoPE.
$s_{u,i}, \alpha_{u,i}$	Attention score and its softmax weight over the full context.
$\mathbf{o}_u$	Attention-head output for query token u.
$\mathcal{I}$	Window of reasoning tokens preceding the current crop request.
$\mathcal{H}, h$	Query heads sharing the fixed KV head, and an individual query-head index.
$\mathbf{q}_{h,u}, \mathbf{k}_{x_j}$	Intent-token query in head h and selected visual key in the parent context.
$a_{h,u,j}$	Saved, full-context attention from intent token u, head h, to visual token j.
b_j	Attention averaged over $u \in \mathcal{I}$ and $h \in \mathcal{H}$, before regional normalization.
π_j	Relevance prior normalized over the N selected tokens in B_t; uniform fallback is $1/N$.
$G_j, \mathbf{v}_j^{\mathrm{new}}$	Value gain and adapted value, $\mathbf{v}_j^{\mathrm{new}} = G_j \mathbf{v}_j^0$.
β, clip	Gain coefficient with $1 \le G_j \le 1 + \beta$, and clipping to $[0, 1]$.

A.4 Algorithm Summary

Algorithm 1 summarizes the inference procedure using the notation of Section 4. The parent observation determines the crop coordinates and area ratio; the source bank supplies clean keys and base values. The mapping ψ resolves selected tokens to their source entries, including when the parent itself was formed by reuse. IVE reads current pre-crop reasoning attention to B_t within O_s and normalizes the aggregated attention over the selected visual tokens in B_t.

Algorithm 1 Adaptive visual KV cache reuse for multi-turn reasoning

Require: Original image I, question Q, frozen MLLM, threshold τ, gain coefficient β
Ensure: Final answer, subject to the model's existing stopping limits

Prefill the initial input and retain its clean visual states and regional mappings.
while the existing inference budget is not exhausted **do**
  Decode until a crop request or final answer; retain attention for up to 64 recent generated tokens.
  if a final answer is emitted **then**
    return the answer
  end if
  if decoding stops without emitting a crop request **then**
    break
  end if
  Read crop region B_t, parent O_s, and current insertion position $\mathbf{p}_{\text{ins}}$.
  Compute $r_t = |B_t|/|\Omega_s|$ and cache availability H_t as in Equation (4).
  if $r_t \leq \tau$ **or** $H_t = 0$ **then**
    Map B_t to the original image and obtain the crop pixels.
    Run visual encoding and visual-token prefill to obtain observation O_t.
    Store its clean keys, base values, positions, and mappings in the source bank.
  else
    Resolve $N \geq 1$ selected tokens to source entries $\psi(j)$ and positions $\mathbf{p}_j^{\text{src}}$.
    Set $\mathbf{\Delta} = \mathbf{p}_{\text{ins}} - \mathbf{p}_1^{\text{src}}$ and $\mathbf{p}_j^{\text{new}} = \mathbf{p}_j^{\text{src}} + \mathbf{\Delta}$.
    for each decoder layer and KV head **do**
      Retrieve clean keys $\bar{\mathbf{k}}_j$ and base values $\mathbf{v}_j^0$, aligned through $\psi(j)$.
      if the reasoning window is empty or regional attention mass is zero or unavailable **then**
        Set $\pi_j \leftarrow 1/N$ for all selected tokens.
      else
        Form π_j from saved attention to B_t in O_s using Equations (6)–(7).
      end if
      for each selected token j **do**
        $\mathbf{k}_j^{\text{new}} \leftarrow \mathbf{R}(\mathbf{p}_j^{\text{new}})\bar{\mathbf{k}}_j$ ▷ PKR
        $G_j \leftarrow 1 + \beta \, \text{clip}(N\pi_j - 1, 0, 1)$
        $\mathbf{v}_j^{\text{new}} \leftarrow G_j \mathbf{v}_j^0$ ▷ IVE
      end for
    end for
    Insert the adapted KV states as observation O_t in the active decoding cache.
    Retain mappings to the same source entries; leave the source bank unchanged.
  end if
  Resume text decoding with the new observation in context.
end while
Apply the model's existing termination rule if no final answer has been emitted.

Behavior under consecutive reuse. Each reuse obtains a new positional offset and relevance prior while retrieving the same unmodified source states when appropriate. The gain bound $1 \leq G_j \leq 1 + \beta$ therefore applies independently at every reuse step. This procedure adapts cached evidence to the current decoding context; it does not claim to reproduce the states that full visual prefill would have computed from the new crop and reasoning prefix.

A.5 Visualization of Intent-Guided Value Enhancement

Figure 8 complements Figure 7 with additional examples of the relevance prior and value gains at selected reuse steps.

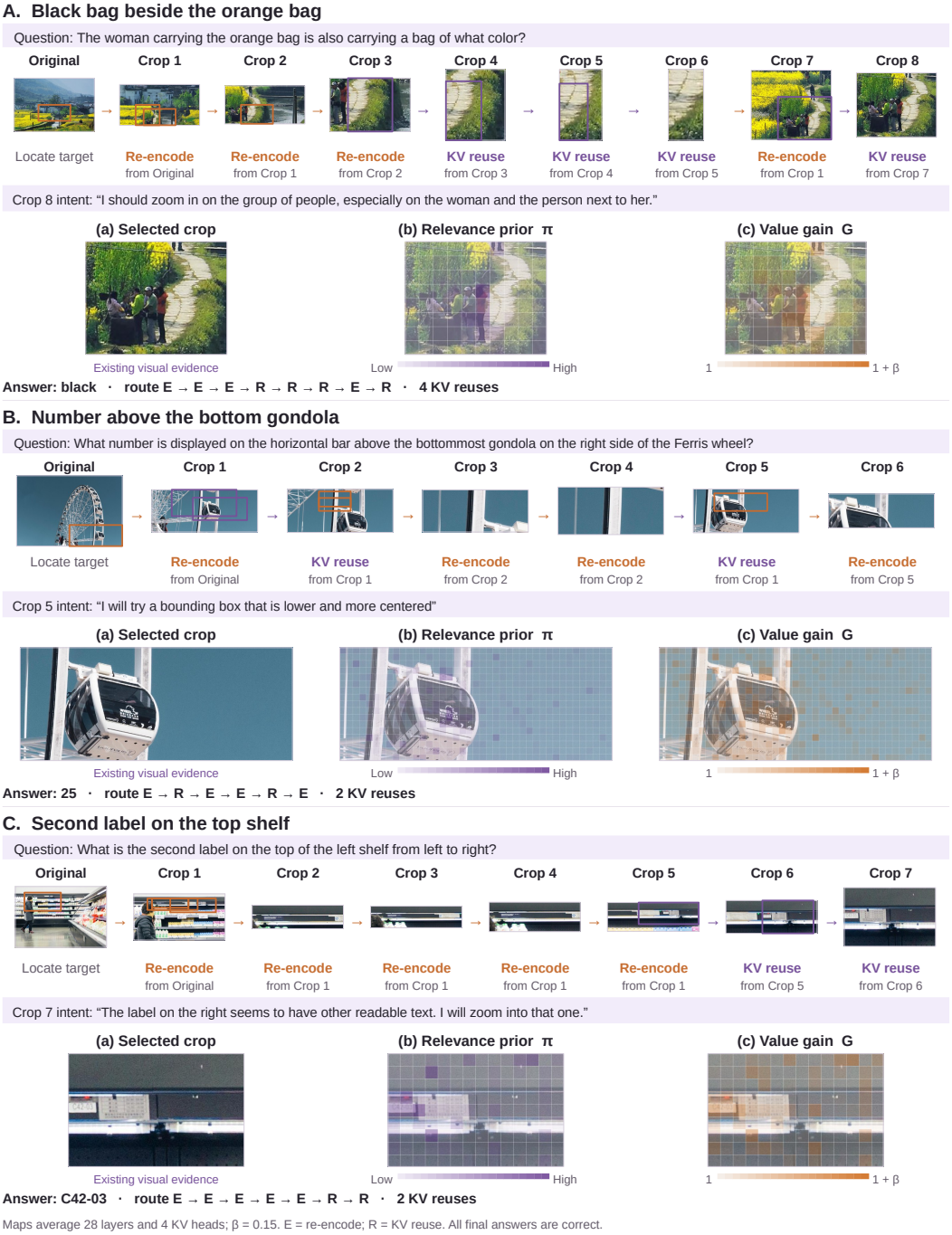


Figure 8: **Additional visualizations of intent-guided value enhancement.** Each example shows a crop trajectory and, for a selected reuse step, (a) the selected crop, (b) the intent-derived relevance prior π, and (c) the value gains G. Orange and purple indicate re-encoding and KV reuse, respectively. Spatial maps average 28 layers and four KV heads, with $\beta = 0.15$.

A.6 Visualization of Multi-Turn KV Cache Reuse

The following examples visualize multi-turn visual KV cache reuse through reasoning excerpts, crop requests, returned observations, and final answers. Re-encoding and reuse steps are marked in each trajectory to show how the model acquires new detail and revisits cached visual evidence.

Original image



Question

what are the words on the orange ship on the right?

Thought 0 (excerpt)
I can see a large orange ship on the right side of the image. I will first isolate that orange ship to get a clearer view of the text on it.

Action 0 → Obs. 1: Re-encoding
grounding([0.6, 0.4, 1.0, 0.7], source="original_image")

Observation 1 · Re-encoding



Thought 1 (excerpt)
In observation 1, I can see an orange ship, and there seems to be some text on it. I will zoom in on the orange ship to read the words.

Action 1 → Obs. 2: KV reuse
grounding([0.4, 0.2, 1.0, 0.9], source="observation_1")

Observation 2 · KV reuse



Thought 2 (excerpt)
In observation 2, I can see the orange ship more clearly. There appears to be some text written on the side of the orange ship. I will zoom into that area to read the text.

Action 2 → Obs. 3: Re-encoding
grounding([0.5, 0.4, 1.0, 0.7], source="observation_2")

Observation 3 · Re-encoding



Thought 3 (excerpt)
In observation 3, I can see some text on the orange ship. It appears to be "FRU INGER". The text is very clear in this observation. I will zoom in one more time to be certain.

Action 3 → Obs. 4: KV reuse
grounding([0.0, 0.0, 0.6, 0.5], source="observation_3")

Observation 4 · KV reuse



Thought 4 (excerpt)
In this final observation, the text on the orange ship is clearly visible. The text says 'FRU INGER'. I am confident in this answer and no more grounding actions are needed.

Action 4 Correct · exact ground truth
final("FRU INGER")

Figure 9: **Reading the name on an orange ship.** Crops 1 and 3 use re-encoding to locate the ship and resolve its lettering; Crops 2 and 4 use KV reuse. The final answer, `FRU INGER`, matches the ground truth.

Original image



Question

What is the first word of the last line on the grey sign on the far right?

Thought 0 (excerpt)
The user wants me to identify the first word on the last line of the grey sign on the far right. I will first locate the sign in the original image and then use the grounding tool to get a closer view.

Action 0 → Obs. 1: Re-encoding
grounding([0.8, 0.5, 1.0, 0.7], source="original_image")

Turn 1 Observation 1
Re-encoding



Thought 1 (excerpt)
I've already cropped the region containing the sign in Observation 1. The sign has several lines of text. I will zoom in on the sign to read it more clearly. The last line appears to be "Balcons". To be certain, I will zoom in further on the sign.

Action 1 → Obs. 2: KV reuse
grounding([0.5, 0.3, 1.0, 1.0], source="observation_1")

Turn 2 Observation 2
KV cache reuse (PKR + IVE)



Thought 2 (excerpt)
The last line seems to be "Balcons". The first word of this last line is "Balcons". The text is quite clear, but to be more certain and to make sure there are no other words, I will zoom in on the last part of the sign which says "Balcons".

Action 2 → Obs. 3: Re-encoding
grounding([0.5, 0.7, 1.0, 1.0], source="observation_2")

Turn 3 Observation 3
Re-encoding



Thought 3 (excerpt)
The previous action with bounding box `[0.5, 0.7, 1.0, 1.0]` on observation_2 resulted in a view of the fence instead of the sign. I need to go back to observation 2 and select a better bounding box to isolate the text on the sign.

Action 3 → Obs. 4: KV reuse
grounding([0.5, 0.3, 1.0, 1.0], source="observation_2")

Turn 4 Observation 4
KV cache reuse (PKR + IVE)

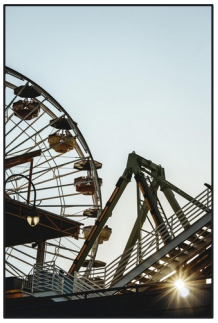


Thought 4 (excerpt)
From observation 4, I can clearly see a grey sign which has multiple lines of text. The last line of text clearly says 'Balcons'. The first word of this line is 'Balcons'. The image is clear enough. No more grounding is needed.

Action 4 Correct · matches ground truth
final("Balcons")

Figure 10: **Reading a sign by revisiting an earlier observation.** After Crop 3 captures the fence, the model returns to Observation 2 for Crop 4 and uses KV reuse. The final answer, `Balcons`, matches the ground truth. Crops 1 and 3 are re-encoded, while Crops 2 and 4 use PKR and IVE.

Original image



Question

What is the white number on the Ferris wheel box in the upper left corner of the picture?

Thought 0 (excerpt)
I need to locate the Ferris wheel and then focus on the box in the upper left. The top left portion of the Ferris wheel contains some gondolas.

Action 0 → Obs. 1: Re-encoding
grounding([0.0, 0.1, 0.5, 0.4], source="original_image")

Observation 1 · Re-encoding



Thought 1 (excerpt)
In Observation 1, I can see a portion of the Ferris wheel. There is a box with some text or numbers on it. I need to zoom in on this specific box to read the number clearly. The box with the number appears to be on the left side of the observation.

Action 1 → Obs. 2: KV reuse
grounding([0.0, 0.5, 0.6, 1.0], source="observation_1")

Observation 2 · KV reuse



Thought 2 (excerpt)
I have zoomed in on the Ferris wheel and I can see a box with a number on it. In observation 2, I can see the number "19" in white on one of the cabins of the Ferris wheel. I will zoom in further to be certain.

Action 2 → Obs. 3: KV reuse
grounding([0.0, 0.5, 0.7, 1.0], source="observation_2")

Observation 3 · KV reuse



Thought 3 (excerpt)
In observation 3, I have successfully zoomed in on one of the gondolas on the Ferris wheel. I can clearly see the number '19' written in white on the side of the gondola. The number is clearly visible and readable. Therefore, I can confidently provide the final answer.

Action 3 Correct · exact ground truth
final("19")

Figure 11: **Reading a gondola number through consecutive reuse.** After one re-encoding step, Crops 2 and 3 reuse cached states to inspect the selected gondola. The model reads and confirms the number `19`.

Original image



Observation 1 · Re-encoding



Observation 2 · Re-encoding



Observation 3 · KV reuse



Observation 4 · KV reuse



Observation 5 · KV reuse



Question

The woman carrying the orange bag is also carrying a bag of what color?

Thought 0 (excerpt)
First, I need to locate the woman with an orange bag. I'll start by cropping the image to focus on this person.

Action 0 → Obs. 1: Re-encoding
grounding([0.3, 0.5, 0.7, 0.9], source="original_image")

Thought 1 (excerpt)
In Observation 1, I can see a group of people on the path. There is a woman in an orange top, and it seems she might be carrying something. I'll need to zoom in further to see what she is carrying and what color the second bag is.

Action 1 → Obs. 2: Re-encoding
grounding([0.3, 0.5, 0.7, 0.9], source="observation_1")

Thought 2 (excerpt)
In Observation 2, I can clearly see a woman in an orange top. However, it is hard to distinguish the color of her other bag due to the image quality and the angle. I'll zoom in on the group of people to get better details.

Action 2 → Obs. 3: KV reuse
grounding([0.0, 0.2, 0.6, 1.0], source="observation_2")

Thought 3 (excerpt)
In observation 3, I can see a woman wearing a bright orange jacket. This woman is carrying a bag that seems to be black or dark-colored. I need a closer view to the woman in the orange jacket to identify the color of her bags more accurately.

Action 3 → Obs. 4: KV reuse
grounding([0.2, 0.2, 0.6, 0.9], source="observation_3")

Thought 4 (excerpt)
On the ground next to her, there is another bag. The color of this second bag seems to be black. However, the image is a bit blurry. I need to take one more closer view to confirm the color of both bags.

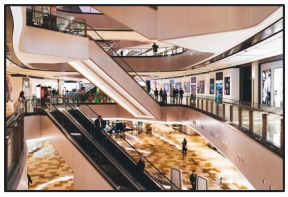
Action 4 → Obs. 5: KV reuse
grounding([0.2, 0.3, 0.7, 1.0], source="observation_4")

Thought 5 (excerpt)
In observation 5, I have a closer view of the group of people. I can see the woman in the orange jacket clearly. She is holding a light-colored bag. Now I can clearly see another bag at her feet, and it is black. The image is clear enough to identify the color.

Action 5 Correct · exact ground truth
final("black")

Figure 12: **Inspecting a small object through three consecutive reuse steps.** Two re-encoding steps locate the group of people. Crops 3–5 then reuse cached states to inspect the woman's other bag, yielding the correct answer, `black`.

Original image



Question

What is the text below "CINKER PICTURES"?

Thought 0 (excerpt)
I need to first locate the text "CINKER PICTURES" in the image. I will zoom in on this area to read the text clearly.

Action 0 → Obs. 1: Re-encoding
grounding([0.5, 0.2, 1.0, 0.6], source="original_image")

Observation 1 · Re-encoding



Thought 1 (excerpt)
In observation 1, I can see a row of posters on the wall, and I can faintly see some text on some of them. I will start by focusing on the posters that seem to have English text on them, located in the middle of observation 1.

Action 1 → Obs. 2: KV reuse
grounding([0.1, 0.3, 0.7, 0.8], source="observation_1")

Observation 2 · KV reuse



Thought 2 (excerpt)
In Observation 2, I see a row of posters on the wall. I need to find the poster that says "CINKER PICTURES". I see a poster on the left side of this cropped image that seems to have that text. I will zoom in on that poster to read the text more clearly.

Action 2 → Obs. 3: Re-encoding
grounding([0.1, 0.3, 0.4, 0.7], source="observation_2")

Observation 3 · Re-encoding



Thought 3 (excerpt)
I can see the text "CINKER PICTURES" on the left side of this cropped image. Below that, it seems to say "SHANGHAI". I need to get a clearer view of this text to be sure. I need to zoom in on that specific area to confirm the text.

Action 3 → Obs. 4: KV reuse
grounding([0.0, 0.4, 0.5, 0.9], source="observation_3")

Observation 4 · KV reuse



Thought 4 (excerpt)
Below that, the text "SHANGHAI" is clearly visible, although partially cut off. However, for a more definitive view and to make sure there isn't any other text, I will slightly adjust the bounding box to better center the text of interest.

Action 4 → Obs. 5: KV reuse
grounding([0.0, 0.2, 1.0, 1.0], source="observation_4")

Observation 5 · KV reuse



Thought 5 (excerpt)
In Observation 5, I see a poster with the text "CINKER PICTURES" partially visible at the top left. Directly below it, the text "SHANGHAI" is clearly visible in smaller font size. I am confident in the answer.

Action 5 Correct · exact ground truth
final("SHANGHAI")

Figure 13: **Combining re-encoding and reuse for text inspection.** Crops 1 and 3 use re-encoding to locate the poster, while Crops 2, 4, and 5 use KV reuse. The final two reuse steps revisit the text below "CINKER PICTURES," yielding the correct answer, `SHANGHAI`.

Original image



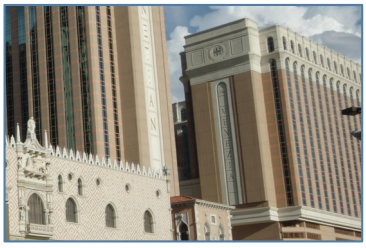
Question

What text is above "VENEZIA" on the building?

Thought 0 (excerpt)
I can see the text "VENEZIA" on the building on the right side of the image. I will zoom in on that area to read the text above it.

Action 0 → Obs. 1: Re-encoding
grounding([0.6, 0.3, 1.0, 0.7], source="original_image")

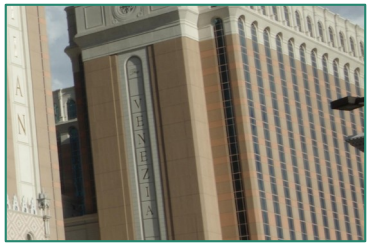
Observation 1 · Re-encoding



Thought 1 (excerpt)
In observation 1, I can see the text "VENEZIA" written vertically on the building. I need to zoom in on the area with the word VENEZIA to read the text written above it.

Action 1 → Obs. 2: KV reuse
grounding([0.4, 0.2, 1.0, 0.8], source="observation_1")

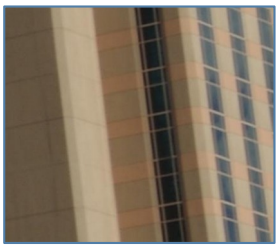
Observation 2 · KV reuse



Thought 2 (excerpt)
In observation 2, I can see the word "VENEZIA" written vertically on the building. The text appears to be "The", but the image quality is not sufficient to be 100% certain. I'll need to zoom in on the text above "VENEZIA" to get a clearer view.

Action 2 → Obs. 3: Re-encoding
grounding([0.45, 0.3, 0.75, 0.7], source="observation_2")

Observation 3 · Re-encoding



Thought 3 (excerpt)
My last action resulted in a very blurry image, and it seems I selected a poor bounding box, missing the target text. I will go back to observation 2, where the text is visible, and try a more precise bounding box to get a clearer look at the text above "VENEZIA".

Action 3 → Obs. 4: KV reuse
grounding([0.3, 0.2, 0.9, 0.75], source="observation_2")

Observation 4 · KV reuse



Thought 4 (excerpt)
In observation 4, I can clearly see the word "VENEZIA" written vertically. Above the word "VENEZIA", I can clearly see the text "The". I am confident in this answer, and no further grounding is needed.

Action 4 Correct · exact ground truth
final("The")

Figure 14: **Recovering the target by reusing an earlier observation.** Crops 1 and 3 use re-encoding, while Crops 2 and 4 use KV reuse. After Crop 3 misses the lettering, Crop 4 returns to Observation 2 and includes the text above "VENEZIA," yielding the correct answer, `The`.

A.7 Limitations and Future Work

Our method reuses existing visual representations and cannot recover details absent from the source cache, making full re-encoding necessary when finer evidence is required. The area-based routing rule and attention-derived intent provide practical signals, but may not fully capture the need for new detail or the current reasoning objective. Moreover, PKR and IVE adapt historical states without reproducing all contextual interactions of full visual prefill. Our evaluation focuses on Mini-o3, and retaining source states and attention information introduces additional storage costs. Future work will explore uncertainty-aware routing, selective state refresh, and memory-efficient cache management, together with evaluation across model architectures, visual tools, and inference settings.