BlockRoute: Per-Step Block-Sparse Cross-Modal Attention for Efficient Multimodal Chain-of-Thought Reasoning

Anonymous ACL submission

Abstract

Multimodal large language models attend over hundreds of visual tokens at every layer during chain-of-thought (CoT) generation, even though each reasoning step often depends on only a subset of the image. We introduce BLOCKROUTE, a structured block-sparse cross-modal attention framework for efficient multimodal reasoning. BLOCKROUTE partitions image tokens into a 4×4 spatial grid and selects the top-k visual blocks to expose through a layer-selective attention mask, preserving spatial locality while reducing cross-modal attention computation. We evaluate several routing strategies on a frozen LLaVA-1.5-7B, including training-free similarity-based routing and a learned per-step router trained on multimodal CoT data. At a 50% image-token keep ratio, BLOCKROUTE substantially reduces analytic cross-modal attention FLOPs while maintaining strong task performance. On ScienceQA-IMG, the learned per-step router matches dense accuracy, reaching 64.5% compared with 64.0% for the dense baseline, while using half of the analytic cross-modal attention FLOPs. A training-free question-embedding router also achieves competitive performance at 63.5%, and BLOCKROUTE outperforms the token-level FASTV baseline under the same visual-token budget. These results suggest that structured, reasoning-conditioned cross-modal sparsity is a promising direction for efficient multimodal chain-of-thought reasoning, offering a simple and interpretable alternative to token-level visual pruning.

1 Introduction

Multimodal large language models (MLLMs) such as LLaVA-1.5 (Liu et al., 2024a), Qwen-VL (Bai et al., 2024), and InternVL (Chen et al., 2024b) solve visual questions by conditioning language generation on a sequence of visual tokens. For a high-resolution image, this visual sequence can contain hundreds of projected patch embeddings. In LLaVA-1.5 at 336^2 resolution, for example, the image contributes 576 visual tokens. During chain-of-thought (CoT) generation, every generated token attends to these visual tokens at every LLM layer. Thus, even a short reasoning trace repeatedly pays the cost of dense cross-modal attention. As illustrated in Fig. 1(a), CoT decoding repeatedly attends to hundreds of visual tokens at every layer, leading to substantial redundant cross-modal computation.

However, as shown in Fig. 1(b), token-level pruning often destroys coherent spatial regions by selecting isolated visual tokens. This computation pattern is stronger than what many reasoning steps appear to require. Empirical studies of MLLMs show that visual attention is often concentrated on a small subset of patches (Chen et al., 2024a), that visual importance changes across decoding depth (Xing et al., 2025), and that many visual tokens contribute little signal at inference (Shang et al., 2024; Zhang et al., 2025). These observations have motivated visual-token reduction methods. Most existing methods first score individual visual tokens, retain a token subset, and continue generation with that selected token sequence. This token-level view is effective for reducing the number of visual embeddings, but it does not explicitly preserve the spatial region structure of the image.

CoT reasoning creates a more specific design requirement. A reasoning step may need to inspect a local region, compare two nearby regions, or maintain a coarse view of the whole image. The relevant visual evidence can therefore vary with the question and with the current reasoning state. The key question is whether cross-modal attention can be made sparse in a way that is structured, interpretable, and compatible with step-wise reasoning. Rather than selecting isolated visual tokens, we would like a routing unit that corresponds to an image region and can be changed as reasoning proceeds.

Text-only sparse-attention research provides a useful precedent. Native Sparse Attention (NSA)

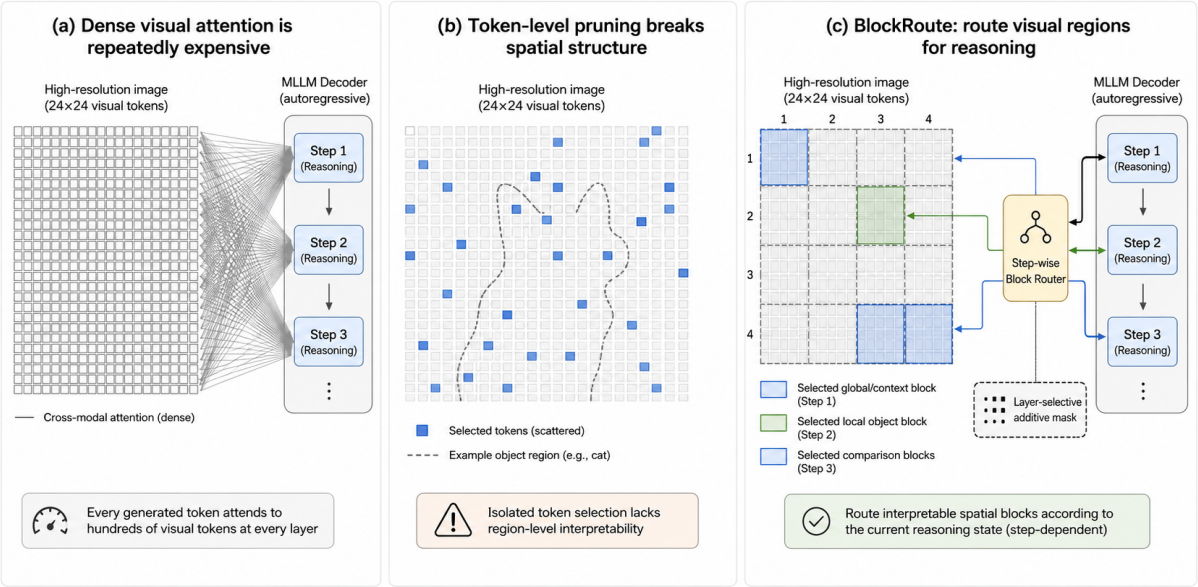

Figure 1: Comparison between dense cross-modal attention, token-level visual pruning, and the proposed BLOCKROUTE framework. (a) Dense cross-modal attention repeatedly attends to hundreds of visual tokens during CoT reasoning. (b) Token-level pruning reduces token count but often fragments coherent spatial regions. (c) BLOCKROUTE performs reasoning-conditioned routing over interpretable spatial blocks, enabling structured sparse cross-modal attention while preserving region-level spatial structure.

(Yuan et al., 2025) and DeepSeek Sparse Attention (DSA) (DeepSeek-AI, 2025) show that sparse attention can be organized through hardware-aligned structure. RaaS (Liu et al., 2025) and ThinKV (Wu et al., 2026) further show that reasoning behavior can inform attention sparsity. However, the corresponding multimodal problem is different because the sparse units should respect the two-dimensional image layout. This motivates structured spatial routing over visual regions instead of isolated tokens (Fig. 1(c)).

We propose BLOCKROUTE, a block-sparse cross-modal attention framework for multimodal CoT reasoning. BLOCKROUTE partitions the 24×24 visual-token grid into a 4×4 grid of spatial blocks, scores these blocks with lightweight routing rules, and exposes only the selected top-k blocks to cross-modal attention through a layer-selective additive mask. This design keeps visual regions intact, gives a fixed and interpretable compute budget, and supports both prompt-level and per-step routing. We evaluate training-free routing rules and a learned per-step router on a frozen LLaVA-1.5-7B, using a token-level FASTV (Chen et al., 2024a) baseline at the same visual-token budget.

Our contributions are three-fold:

- We formulate structured block-sparse cross-modal attention for MLLMs, using a 2D visual block grid and a layer-selective additive mask that preserves region-level spatial structure.
- We instantiate BLOCKROUTE with training-free routers and a learned per-step router, providing a simple interface for prompt-conditioned and reasoning-conditioned visual block selection.
- On ScienceQA-IMG, BLOCKROUTE matches dense accuracy at a 50% visual-token keep ratio, outperforms token-level FASTV under the same budget, and supports layer-placement and grid-granularity analysis.

Overall, BLOCKROUTE shows that structured cross-modal sparsity can preserve dense multimodal reasoning quality while substantially reducing the analytic visual-attention budget.

2 Related Work

Visual Token Reduction for MLLMs. A large body of work reduces the number of visual tokens an MLLM has to process. FASTV (Chen et al., 2024a) drops low-attention visual tokens after layer 2 of the LLM. LLaVA-PruMerge (Shang et al., 2024) prunes and merges with class-attention from the visual encoder. SparseVLM (Zhang et al., 2025) uses text-guided per-layer rank-adaptive sparsification. Recent variants include VisionZip (Wang

et al., 2025b), HiRED (Arif et al., 2024), VTW (Lin et al., 2025b), PyramidDrop (Xing et al., 2025), FitPrune (Ye et al., 2025), ATP-LLaVA (Ye et al., 2024), ZipVL (He et al., 2024), MustDrop (Liu et al., 2024b), DyCoke (Tao et al., 2025), and DivPrune (Alvar et al., 2025). Most of these are token-level methods: they produce a retained token set per image and typically operate without an explicit region-level routing interface. Our method differs by retaining spatial blocks and selecting a different block subset per reasoning step.

Structured and Reasoning-Aware Attention Sparsity. Sparse-attention designs span fixed-pattern (Zaheer et al., 2020; Beltagy et al., 2020), content-based (Kitaev et al., 2020; Roy et al., 2021), and expert-routed (Fedus et al., 2022; Lin et al., 2025a) families. Recent text-side systems such as NSA (Yuan et al., 2025) and DSA (DeepSeek-AI, 2025) combine coarse compression with fine selection in hardware-aligned sparse-attention mechanisms. RaaS (Liu et al., 2025) shows that long-decode reasoning has a distinctive attention pattern that admits sparse attention. ThinKV (Wu et al., 2026) classifies reasoning thoughts into reasoning, execution, and transition types and evicts KV-cache by sparsity per thought type. Together, these works show that sparsity can be structured and reasoning-aware, but they focus on text-side attention rather than cross-modal visual grounding. BLOCKROUTE extends this principle to multimodal CoT by routing cross-modal attention over spatial image blocks.

Multimodal Region and Grid Selection. Visual CoT (Shao et al., 2024) generates bounding boxes inside the CoT and re-encodes crops. V* (Wu and Xie, 2024) wraps the MLLM in a visual-search loop. VLM-R3 (Jiang et al., 2025) trains region-conditioned RL to crop and re-inject sub-images. Chain-of-Focus (Cao et al., 2025), VGR (Lu et al., 2025), and Qwen-LookAgain (Chen et al., 2025) are similar tool-loop or token-route approaches. TokenPacker (Li et al., 2025) compresses ViT tokens via point-to-region injection. Cambrian-1 (Tong et al., 2024) introduces a spatial vision aggregator. MQT-LLaVA (Hu et al., 2024), LLaVA-SP (Lou et al., 2025), and FOLDER (Wang et al., 2025a) are related. These works motivate region-level and grid-aware visual representations. Compared with them, BLOCKROUTE is attention-side, operates throughout CoT generation, and uses the 2D grid as the unit of block-sparse cross-modal attention.

3 Method: BLOCKROUTE

3.1 Overview

Let an MLLM produce a chain-of-thought answer $y_{1:T}$ given a text prompt x and image I. The image is encoded by a vision tower into an $H \times W$ patch grid (for LLaVA-1.5: $24 \times 24 = 576$ patches), each projected into the LLM hidden space and prepended to the text. At every LLM layer, every text query attends over all 576 visual keys, regardless of whether the current reasoning step needs them.

BLOCKROUTE changes the unit of cross-modal sparsity from individual visual tokens to spatial visual blocks. The method has three stages. First, the projected visual tokens are grouped into a fixed $B \times B$ grid. Second, a router assigns a score to each block and selects a top-k subset $S_t \subset \{1, \ldots, G\}$, where $G = B^2$. Third, the selected block set is converted into an attention mask, so text queries can attend to visual keys inside S_t while other visual keys are hidden. This pipeline separates the semantic decision of which image regions to expose from the mechanical operation of masking attention logits.

The default configuration uses $B = 4$, giving $G = 16$ blocks of 36 visual tokens each. At the 50%-keep setting, BLOCKROUTE keeps $k = 8$ blocks and therefore exposes exactly 288 of 576 visual tokens. Because each block corresponds to a contiguous image region, the selected set is easy to visualize and the compute budget is determined directly by k. We instantiate routing rules ranging from training-free per-prompt heuristics to a learned per-step router.

3.2 Block Grid

Define the token-to-block map $\pi : \{1, \ldots, 576\} \to \{1, \ldots, G\}$ by tiling the 24×24 patch grid into $B \times B$ axis-aligned blocks. Each block contains $576/G$ contiguous patches. For the default $B = 4$, the visual grid is divided into 16 blocks, and each block contains a 6×6 patch neighborhood. This choice is coarse enough to keep routing low-dimensional and fine enough to distinguish image regions such as top, bottom, left, right, and center areas. The same definition naturally extends to 2×2 or 8×8 grids, which we evaluate in the granularity ablation. We additionally define a fixed token-level keep mask

$$m_t[i] = \mathbb{I}[\pi(i) \in S_t], \quad i = 1, \ldots, 576, \qquad (1)$$

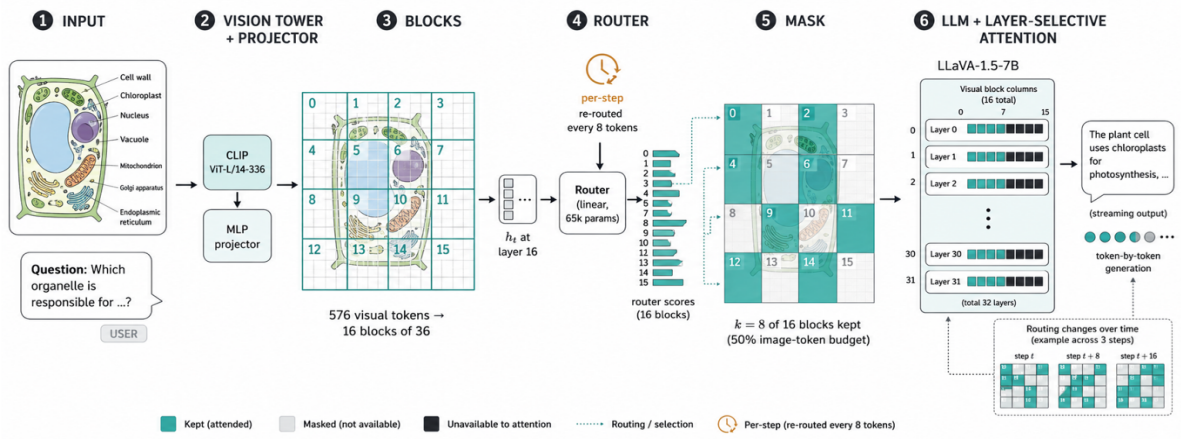

Figure 2: BLOCKROUTE pipeline. The 24×24 patch grid (576 visual tokens) is partitioned into a 4×4 block grid (16 blocks of 36 tokens). A router takes the LLM hidden state $h_t^{\ell_{\text{router}}}$ at layer 16 and predicts a top-k block subset; cross-modal attention is masked to that subset at every LLM layer via a layer-selective additive bias. The router can be training-free (CLIP-SIM, Q-COSINE) or a trained linear head (ROUTER, ROUTER_STEP); the per-step variant re-routes every r generated tokens.

that turns the block decision S_t into a per-image-token visibility mask. The mask is binary at the token level, but the routing decision is made at the block level. This distinction is important: BLOCKROUTE still interfaces with a standard transformer attention implementation through token positions, while the router operates over a compact and interpretable spatial representation.

3.3 Routing Strategies

The router maps the current prompt or reasoning state to a score vector in $\mathbb{R}^G$. The top-k entries of this vector define S_t. We use several routers to separate the effect of the block-sparse attention pattern from the effect of router training. The training-free routers test whether simple similarity signals are enough to choose useful image regions. The learned routers test whether the current LLM hidden state can support reasoning-conditioned routing. We evaluate four routing rules, ordered from training-free to learned, plus a non-block-structured token-level baseline:

- **Random:** pick k blocks uniformly at random; serves as the no-information control.
- **CLIP-SIM:** score block b by $s_b = \text{mean}_{i \in b} \cos(P_i, \bar{e}_x)$, where P_i is the multimodal-projected vision feature for patch i and $\bar{e}_x$ is the mean CLIP-text feature of the question x. Pick top-k.
- **Q-COSINE:** score block b by cosine similarity between the block centroid and the mean of LLM input-token embeddings of the question.
- **ROUTER:** a linear head $W \in \mathbb{R}^{D \times G}$ takes the LLM hidden state at layer $\ell_{\text{router}} = 16$ at the prompt-end position and emits per-block scores. The head is trained on 200 examples from M3CoT via teacher-distilled CoTs plus a load-balance loss (Fedus et al., 2022) and a KL distillation term. Only the linear head is trained (65,552 parameters); the LLaVA backbone is frozen.
- **ROUTER_STEP:** the same trained linear head, re-queried every $r = 8$ generated tokens at the current LLM hidden state, producing a fresh top-k block decision.
- **FASTV baseline:** score image tokens by mean text-to-image attention at LLM layer 2 from a single forward pass and keep the top $36k$ tokens to match BLOCKROUTE's image-token budget.

The single-shot routers select blocks once from the prompt representation and reuse the same block set during generation. The per-step router recomputes block scores every r generated tokens. This gives the model a lightweight mechanism to adjust the visible image regions as the CoT evolves, while keeping the backbone frozen and adding only a small linear head. The comparison to FASTV is budget-matched: for a given k, BLOCKROUTE exposes $36k$ visual tokens, and FASTV keeps the same number of individual tokens.

3.4 Layer-Selective Masking

We can apply the block mask at any subset $\mathcal{L} \subseteq \{0, \ldots, L-1\}$ of LLM layers (default $L = 32$). At a layer $\ell \in \mathcal{L}$, BLOCKROUTE sets the attention logits on masked image keys to a large negative value:

$$\tilde{A}_{ij} = A_{ij} - M \cdot \mathbb{I}[\pi(j-s) \notin S_t], \quad s \leq j < e, \tag{2}$$

where $A_{ij} = q_i^\top k_j / \sqrt{d}$ and $[s, e)$ is the image-token span in the input sequence. We use `torch.finfo(dtype).min` and assign rather than add, to avoid fp16 underflow when combining with the existing causal mask. For layers outside $\mathcal{L}$, attention is left unchanged. This lets us study where cross-modal sparsity is most useful: all-layer masking gives the largest analytic reduction, while partial-layer masking exposes how early, middle, or late layers react to sparse visual context. In the main setting, we apply the mask to all 32 LLM layers for a clean comparison of routing strategies.

Compute Accounting. Our analytic cross-modal attention budget is proportional to the number of visible visual keys. Dense attention exposes all 576 keys. With a 4×4 grid and $k = 8$, BLOCKROUTE exposes $8 \times 36 = 288$ keys, so the relative cross-modal attention budget is 0.50. This accounting is independent of the router type, which allows random, training-free, learned, and token-level baselines to be compared under the same visual-token budget.

4 Experiments

4.1 Setup

Backbone. We use LLaVA-1.5-7B (Liu et al., 2024a): a Vicuna-7B-v1.5 LLM (Chiang et al., 2023), CLIP ViT-L/14-336 (Radford et al., 2021), and a two-layer MLP projector. The model has 32 LLM layers, hidden dimension 4096, 32 heads, and 576 image tokens. All experiments keep the LLaVA backbone frozen. For learned routing variants, only a lightweight linear router head is trained, so differences between dense, training-free, and learned settings primarily reflect the visual-routing policy rather than changes to the underlying MLLM.

Benchmarks. We evaluate on ScienceQA-IMG (Lu et al., 2022), using image-bearing validation examples. The main sweep uses $n = 200$ examples to compare dense attention, token-level pruning, training-free block routing, and learned block routing under the same visual-token budget. Layer-placement and grid-granularity ablations use $n = 60$ examples, and routing-behavior analysis uses $n = 100$ examples. All subsets are sampled with a fixed seed of 0 unless otherwise specified. This setup lets us evaluate both the main accuracy–compute tradeoff and the design choices that determine where and how block sparsity should be applied.

Generation and Evaluation. We use greedy decoding with repetition penalty 1.15 and maximum new tokens 192. Prompts contain the image, question, and answer choices, followed by the instruction: "Think step-by-step, then write 'Answer: X' on a new line." Accuracy is computed from the final extracted answer choice. The same prompts, data order, decoding settings, and answer-extraction script are used for all methods. For the random-routing baseline, we fix the routing seed to make the result reproducible.

Configurations and Fair Comparison. The default block grid is 4×4 ($G = 16$ blocks of 36 tokens). The default layer placement applies the BLOCKROUTE mask at all 32 LLM layers. The default routing layer for learned and cosine routers is layer 16. For the main comparison, all sparse methods use a 50% image-token keep ratio. In BLOCKROUTE, this means selecting $k = 8$ of 16 blocks, exposing exactly 288 of 576 visual tokens. For the token-level FASTV baseline, we keep the top 288 individual visual tokens according to its layer-2 attention score. Thus, BLOCKROUTE and FASTV are compared at the same analytic cross-modal attention budget, while differing only in whether the retained evidence is structured as contiguous spatial blocks or scattered individual tokens.

Router Training. The learned router is a single linear head trained for 200 steps on 200 image-bearing validation examples from M3CoT, sampled with seed 42. We use AdamW with learning rate 2×10^{-4}, weight decay 0, load-balance weight $\lambda_{\text{lb}} = 0.05$, and KL weight $\lambda_{\text{kl}} = 0.5$. Soft routing is used during training and hard top-k routing at evaluation. Only the 65,552-parameter linear head is trained and the LLaVA backbone remains frozen. The same trained head is used for both ROUTER and ROUTER_STEP, so the comparison between them isolates the effect of re-routing during generation.

Compute Accounting. Inference is fp16 with PyTorch SDPA attention; eager attention is used only when scoring FASTV's text-to-image attention map and during router training. Experiments run on a single RTX PRO 6000 (96GB). We report analytic cross-modal attention FLOPs rather than wall-clock latency, because the current implementation applies sparsity through an additive attention mask on top of standard attention kernels. At 50% keep, the number of visible visual keys is reduced from 576 to 288, giving a relative cross-modal attention budget of 0.50. This metric isolates the routing policy and represents the computation that a fused block-sparse implementation would remove.

4.2 Main Results

Table 1 reports ScienceQA-IMG accuracy, image-token budget, and analytic cross-modal FLOPs. All routed methods see exactly 50% of the image tokens, so the comparison isolates the routing decision itself: the methods differ only in which visual evidence they expose and whether the decision is made once or updated during generation. This controlled setting is important because a token budget alone does not determine performance. Two methods can keep the same number of visual tokens while presenting very different spatial evidence to the language model. BLOCKROUTE therefore evaluates not only compression strength, but also the structure of the retained visual context.

ScienceQA-IMG. The trained per-step router matches dense accuracy (64.5 vs. 64.0) while exposing only half of the image tokens to cross-modal attention. The training-free Q-COSINE rule reaches 63.5, matching the trained single-shot router within the same accuracy band and outperforming the FASTV token-level baseline. These results show three consistent trends. First, dense cross-modal attention is not necessary for maintaining ScienceQA-IMG performance in this setting: a 50% block budget is enough for the strongest routed variant to stay at the dense level. Second, structured block selection compares favorably with token-level pruning under the same budget. FASTV keeps 288 individual visual tokens and reaches 61.5%, while Q-COSINE and ROUTER_STEP keep 288 tokens as eight contiguous spatial blocks and reach 63.5% and 64.5%, respectively. This suggests that preserving local image neighborhoods is useful for multimodal reasoning, even when the total number of visible tokens is fixed. Third, step-conditioned routing provides the strongest result among the evaluated variants. The single-shot learned router reaches 62.5%, while re-querying the same lightweight router during generation reaches 64.5%. This supports the central motivation of BLOCKROUTE: different parts of a reasoning trace can benefit from exposing different image regions.

The training-free results are also informative. CLIP-SIM and Q-COSINE use no task-specific router training, yet Q-COSINE remains within 0.5 points of dense. This indicates that useful block choices can already be obtained from the geometry of the projected visual and textual representations. At the same time, the learned per-step router gives a simple path to improve routing without modifying the MLLM backbone. Overall, Table 1 positions BLOCKROUTE as a practical middle ground between dense visual attention and unstructured token pruning: it preserves spatial locality, keeps the compute budget explicit, and can be driven by either training-free similarity or a small learned head.

4.3 Sparsity Sweep

Figure 3 sweeps the keep ratio. On ScienceQA-IMG, the relationship between sparsity and accuracy is encouragingly flat: dense, Q-COSINE, and both trained-router variants stay within the dense ± 1.5 point band from 25% to 75% keep, with ROUTER_STEP at $k = 8$ matching the dense-accuracy band at half the analytic cross-modal FLOPs.

4.4 Layer Placement

Table 2 sweeps the placement set $\mathcal{L}$ at $k = 8$ on ScienceQA-IMG ($n = 60$). Placement provides a useful control knob. Masking the last 16 layers preserves accuracy as well or better than dense, while the every-four-layer placement also remains competitive with only eight masked layers. This is consistent with the FASTV-style finding that visual signal is especially important in early layers and suggests that later-layer block routing is an attractive operating point.

4.5 Block-Grid Granularity

Table 3 compares grids at matched 50% keep on ScienceQA-IMG. The default 4×4 grid gives 16 routing decisions per prompt and provides a practical balance between coarse (2×2) and fine (8×8)

Method	Acc (%)	vis. tok.	xmod-FLOPs
Dense (baseline)	64.0	576	1.00
Random ($k = 8$)	60.0	288	0.50
FASTV (token-level, layer-2 score)	61.5	288	0.50
CLIP-SIM (training-free)	59.5	288	0.50
Q-COSINE (training-free)	63.5	288	0.50
ROUTER (trained, single-shot)	62.5	288	0.50
ROUTER_STEP (trained, per-step)	**64.5**	288	0.50

Table 1: ScienceQA-IMG main results at 50% image-token keep ratio (LLaVA-1.5-7B, all 32 layers masked). All non-dense methods see exactly 288/576 image tokens. ROUTER and ROUTER_STEP share a single linear head trained on M3CoT. Cross-modal attention FLOPs are analytic relative to dense.

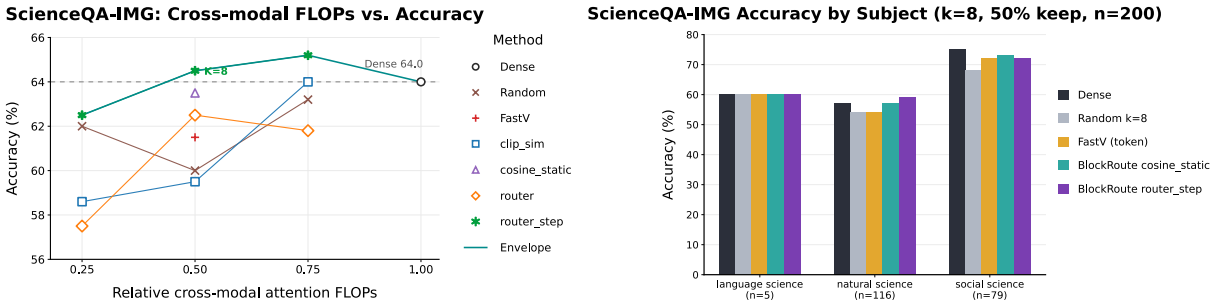

Figure 3: (left)Cross-modal attention FLOPs vs. ScienceQA-IMG accuracy. Each marker is one (method, k) configuration; the dashed teal line traces the empirical envelope across the swept methods. The dotted horizontal line is dense accuracy. ROUTER_STEP ($k = 8$, star) matches the dense-accuracy band at half the analytic cross-modal FLOPs.(right) ScienceQA-IMG accuracy by subject ($k = 8$, 50% keep, $n = 200$). BlockRoute variants track dense performance across subject groups, with strong results on social-science and natural-science examples.

Layers $\mathcal{L}$	$\lvert\mathcal{L}\rvert$	CLIP-SIM	Q-COSINE
all (0–31)	32	50.0	58.3
first16 (0–15)	16	48.3	56.7
last16 (16–31)	16	**58.3**	**58.3**
mid8 (12–19)	8	53.3	55.0
every4	8	56.7	56.7

Table 2: Layer placement on ScienceQA-IMG ($k = 8/16$, $n = 60$). Dense at this n is 56.7%. "every4" denotes layers $0, 4, 8, \ldots, 28$. Higher is better.

Grid	Tok./blk	Rand.	CLIP-SIM
2×2	144	50.0	51.7
4×4	36	60.0	59.5
8×8	9	**55.0**	51.7

Table 3: Block-grid granularity on ScienceQA-IMG, with 50% blocks kept and all 32 layers masked. The corresponding (G, k) values are $(4, 2)$, $(16, 8)$, and $(64, 32)$. 2×2 and 8×8 rows are on $n = 60$; the 4×4 row is on $n = 200$. Dense at $n = 60$ is 56.7%.

grids.

Analytic Compute Reduction. By construction, BLOCKROUTE reduces the number of visual keys visible to cross-modal attention from 576 to 288 at the default 50%-keep setting. This corresponds to a 0.50 relative cross-modal attention FLOP budget in our analytic accounting, while retaining spatially contiguous image regions rather than isolated tokens. Because the routing unit is a block rather than an individual token, the compute budget is easy to reason about: keeping k of 16 blocks exposes exactly $36k$ visual tokens. This makes the method convenient for controlled comparisons against token-level pruning baselines and for deployment settings where the visual-attention budget must be fixed in advance.

5 Analysis

5.1 Where Does Block Routing Help?

The ScienceQA-IMG results suggest that block routing is especially effective when questions can be answered from broad, semantically coherent visual regions. The 4×4 grid keeps local neighborhoods intact, so routing decisions preserve region-level structure while reducing the number of visual tokens exposed to the LLM. This design contrasts with token-level pruning, where selected visual tokens may be spatially scattered and harder to in-

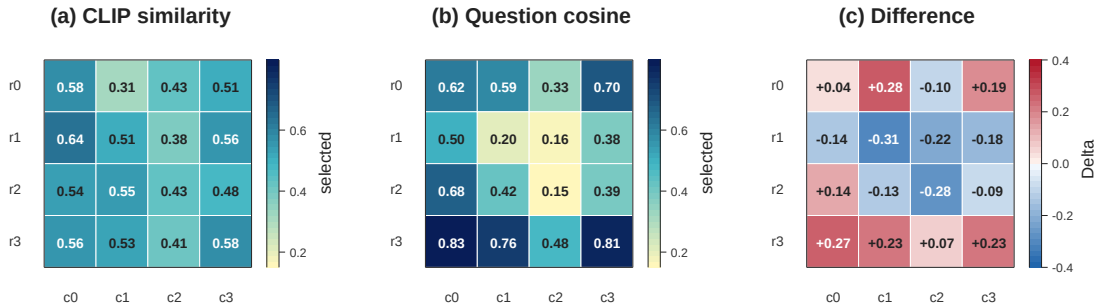

Figure 4: Per-block selection frequency for the two training-free routers (ScienceQA-IMG, $n = 100$, $k = 8$). (a) CLIP similarity is broadly distributed. (b) Question-embedding cosine produces a more concentrated spatial pattern. (c) Difference between the two selection-frequency maps.

terpret. The region-level interface is also useful for analysis: a selected block corresponds to an explicit part of the image grid, so route choices can be inspected, averaged, or compared across examples. This gives BLOCKROUTE a lightweight interpretability advantage over unstructured token dropping, where the retained set is often difficult to summarize beyond a binary token mask.

Subject-Level Breakdown. Figure 3 breaks ScienceQA-IMG accuracy down by subject area. The results show that BLOCKROUTE variants closely follow dense performance across the sampled subject groups. On social-science examples, structured block routing preserves the dense accuracy band while using half of the visual tokens. On natural-science examples, Q-COSINE and ROUTER_STEP remain competitive, indicating that block-level routing can retain useful visual evidence across visually grounded question types. Language science has only five examples in the sampled subset and is included for completeness. These subject-level results complement the aggregate scores in Table 1: the gains are not driven by a single subject category, but arise from a consistent ability to retain enough visual context for multiple kinds of ScienceQA-IMG questions.

5.2 Routing Behavior

We profile the two training-free routers on $n = 100$ ScienceQA-IMG items at $k = 8$. Figure 4 shows that CLIP similarity distributes selections broadly across the image grid, while question-embedding cosine produces a more concentrated spatial pattern. The two strategies pick different blocks on the same items yet both provide competitive accuracy, suggesting that BLOCKROUTE supports multiple viable routing styles. This flexibility is useful: applications can choose between broad coverage and sharper spatial selectivity without changing the underlying model architecture. The block-frequency view also illustrates why a structured routing space is attractive. Instead of reporting a long list of retained patch indices, the routing behavior can be summarized as a small 4×4 heatmap. This exposes whether a method behaves like a broad-coverage selector, a spatially concentrated selector, or a hybrid of the two. In practice, this makes BLOCKROUTE easy to audit: the same routing outputs used for masking can be reused directly for visualizations, per-subject analysis, and comparisons between training-free and learned routers.

5.3 Discussion

Our results suggest that multimodal CoT reasoning does not require uniformly dense cross-modal attention throughout decoding. Instead, the relevant visual evidence is often sparse, structured, and reasoning-dependent, making spatial block routing a natural alternative to token-level pruning. Although BLOCKROUTE is implemented as a lightweight masking mechanism on top of a frozen LLaVA-1.5 backbone, the same design could naturally extend to larger multimodal models and more adaptive routing policies.

6 Conclusion

We introduced BLOCKROUTE, a structured block-sparse cross-modal attention framework for multimodal chain-of-thought reasoning. By routing attention over interpretable spatial blocks instead of isolated visual tokens, BLOCKROUTE substantially reduces analytic cross-modal attention cost while preserving reasoning performance on ScienceQA-IMG. Our results suggest that reasoning-conditioned spatial routing is a promising direction for efficient multimodal reasoning.

Limitations

This study focuses on LLaVA-1.5-7B at 336^2 resolution and uses ScienceQA-IMG as the main benchmark for the reported results. Future work should evaluate the same block-routing pattern on larger and newer MLLMs, higher-resolution image encoders, and broader multimodal reasoning suites. The current implementation applies routing through an additive attention mask, so the reported compute reduction is analytic rather than a fused-kernel wall-clock speedup. A dedicated block-sparse attention kernel would be the natural systems extension. Finally, the learned router is trained with a compact set of 200 examples from M3CoT; scaling router supervision and jointly training routing with the backbone may further improve robustness across domains.

References

Saeed Ranjbar Alvar and 1 others. 2025. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kazi Hasan Arif and 1 others. 2024. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models. In *Advances in Neural Information Processing Systems*.

Jinze Bai and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Yi Cao and 1 others. 2025. Chain-of-focus: Adaptive visual search and zooming for multimodal reasoning. *arXiv preprint arXiv:2505.15436*.

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

Yujie Chen and 1 others. 2025. Qwen-lookagain: Look again, think twice — visual token copy and route for reflective reasoning. *arXiv preprint arXiv:2505.23558*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, Ion Stoica, and Eric P Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90% chatgpt quality. https://lmsys.org/blog/2023-03-30-vicuna/.

DeepSeek-AI. 2025. Deepseek-v3.2: Hierarchical sparse attention with lightning indexer. *arXiv preprint arXiv:2512.02556*.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*.

Yefei He and 1 others. 2024. Zipvl: Efficient large vision-language models with dynamic token sparsification and kv cache compression. *arXiv preprint arXiv:2410.08584*.

Wenbo Hu, Zi-Yi Dou, Liunian Harold Li, Amita Kamath, Nanyun Peng, and Kai-Wei Chang. 2024. Matryoshka query transformer for large vision-language models. *arXiv preprint arXiv:2405.19315*.

Chao Jiang and 1 others. 2025. Vlm-r3: Region recognition, reasoning, and refinement for enhanced multimodal chain-of-thought. In *Advances in Neural Information Processing Systems*.

Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. 2020. Reformer: The efficient transformer. In *International Conference on Learning Representations*.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jianke Qin, Jianbing Zhu, and Lei Zhang. 2025. Tokenpacker: Efficient visual projector for multimodal llm. *International Journal of Computer Vision*.

Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Junwu Huang, Junjie Zhang, Yatian Pang, Munan Ning, and Li Yuan. 2025a. Moe-llava: Mixture of experts for large vision-language models. *IEEE Transactions on Multimedia*.

Zhihang Lin and 1 others. 2025b. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Junhao Liu and 1 others. 2025. Raas: Reasoning-aware attention sparsity for efficient long-decode language models. In *Findings of the Association for Computational Linguistics*.

Ting Liu and 1 others. 2024b. Mustdrop: Multi-stage visual token dropping for efficient mllm inference. *arXiv preprint arXiv:2411.10803*.

Hao Lou and 1 others. 2025. Llava-sp: Spatial tokens for multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*.

Xiao Lu and 1 others. 2025. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. 2021. Efficient content-based sparse attention with routing transformers. *Transactions of the Association for Computational Linguistics*, 9:53–68.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. 2024. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*.

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

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. 2025. Dycoke: Dynamic compression of tokens for fast video large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, and 1 others. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems*.

Haicheng Wang and 1 others. 2025a. Folder: Accelerating multi-modal large language models with enhanced performance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Senqiao Wang, Yukang Chen, Sitong Zhang, Zhuotao Liu, Yulu Xu, Hengshuang Zhao, Chen Change Loy, Dahua Lin, and Jiaya Jia. 2025b. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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

Yuxin Wu and 1 others. 2026. Thinkv: Thought-adaptive kv-cache compression for reasoning models. In *International Conference on Learning Representations*.

Long Xing and 1 others. 2025. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Xubing Ye, Yukang Gan, Xiaoke Huang, Yixiao Ge, Ying Shan, and Yansong Tang. 2025. Fitprune: Fast and statistically optimal visual token pruning for large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Xubing Ye and 1 others. 2024. Atp-llava: Adaptive token pruning for large vision language models. *arXiv preprint arXiv:2412.00447*.

Jingyang Yuan and 1 others. 2025. Native sparse attention: Hardware-aligned and natively trainable sparse attention. *arXiv preprint arXiv:2502.11089*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and 1 others. 2020. Big bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. 2025. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *Proceedings of the International Conference on Machine Learning*.