# WALKPRUNER: COGNITIVELY INSPIRED VISUAL TOKEN PRUNING FOR SPATIAL UNDERSTANDING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Visual token pruning reduces the computational cost of spatial understanding in multimodal large language models. Multi-view inputs contain repeated scene content, but viewpoint changes reveal complementary observations needed to distinguish spatial relationships. Objects overlapping in one view may be separated in another. Feature-guided pruning can preserve the main scene content while discarding spatial details and contextual cues. Geometry-guided pruning maintains diversity across voxels, yet may concentrate tokens in some viewpoints while overcompressing observations from others. Effective pruning must retain representative scene content together with complementary cues along the observation path. Inspired by non-selective global perception and selective processing in human spatial exploration, we introduce **`WaLKPruner`**, a training-free framework following the principle of ***W**alk **a**nd **L**ook for **K**ey Anchors*. Scene abstraction and viewpoint tracking jointly guide token selection, maintaining a compact scene overview while incorporating complementary observations along the ordered observation path. Constrained selective updates limit changes to the initial scene representation while maintaining a fixed token budget. Experiments on three benchmarks show strong performance across token budgets. Efficiency analyses indicate a favorable balance between computational cost and spatial understanding. Quantitative and qualitative analyses reveal balanced token allocation across viewpoints, while ablations support combining scene and path information and retaining observation order. Code and scripts will be released upon acceptance.

## 1 INTRODUCTION

Visual token pruning reduces the computational cost of spatial understanding (Li et al., 2026a). For multimodal large language models (MLLMs) (Bai et al., 2025; Chen et al., 2024c), multi-view inputs contain substantial redundancy from repeated observations of the same objects and regions. Yet changes in camera position and orientation can reveal previously hidden objects and provide complementary cues about relative positions, occlusion, and scene layout (Wang et al., 2025b). Effective pruning must therefore balance two objectives: representing the scene compactly and preserving viewpoint-complementary evidence needed to distinguish spatial relationships.

Existing approaches broadly follow two pruning strategies, as illustrated in Figure 1. **Feature-guided pruning** methods (Kim et al., 2026; Shen et al., 2026; Fu et al., 2025) select or merge tokens based on importance scores and feature redundancy. Such compression can preserve the main scene content but discard details and contextual cues needed to distinguish spatial relationships. **Geometry-guided pruning** (Li et al., 2026a) uses reconstructed 3D geometry to select tokens and maintain spatial diversity across voxels, incurring reconstruction overhead. It may also retain many tokens from some viewpoints while overcompressing complementary observations from others, weakening the evidence for spatial understanding. Objects overlapping in one view may be separated in another. These limitations motivate a strategy that preserves representative scene content and selectively refines retained evidence using viewpoint changes along the observation sequence.

Human spatial exploration combines non-selective global perception with selective processing of informative regions under limited attentional capacity (Wolfe et al., 2011; Li et al., 2026b). For an observer traversing a scene, this perspective suggests maintaining a scene overview while capturing complementary cues as the viewpoint changes. Following the principle of ***W**alk **a**nd **L**ook for **K**ey*



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.

*Anchors*, we introduce **WaLKPruner**, a training-free framework for cognitively inspired pruning. Given an ordered trajectory, scene abstraction and viewpoint tracking jointly update retained tokens to balance global representation and viewpoint diversity under a fixed budget.

Specifically, scene abstraction selects representative tokens through feature similarity, forming a compact overview of the observed scene. Using predicted camera poses (Wang et al., 2025b), viewpoint tracking accumulates translation and rotation in sequence order to distinguish observation stages, including revisits to similar poses. Building on both, joint scoring combines scene and path coverage to identify candidates for refining the initial selection. Finally, selective update prioritizes candidates from viewpoints with larger local pose changes and replaces retained tokens under constraints on changes to the initial set and scene coverage loss. In Figure 2, WaLKPruner uses viewpoint changes along the observation path to retain more widely distributed observations of the chairs and bookshelf, providing spatial references for the turning decision.

WaLKPruner achieves strong performance on VSI-Bench (Yang et al., 2025a), ScanQA (Azuma et al., 2022), and SQA3D (Ma et al., 2022) across different token budgets. It also offers a favorable balance between computational cost and spatial understanding. Quantitative and qualitative analyses reveal more balanced token allocation across viewpoints and illustrate how complementary observations provide spatial references along the observation path, even at nearby positions with different viewing directions. Ablation studies further support the complementary roles of scene abstraction and viewpoint tracking, as well as the value of preserving observation order.

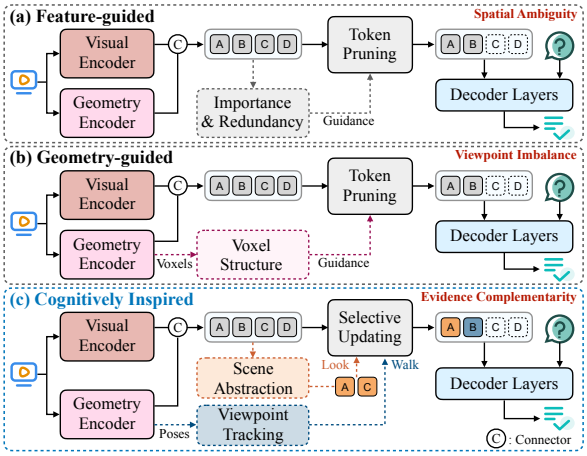


Figure 1: **Comparison of token pruning paradigms** for spatial understanding. **(a)** Feature-guided pruning may obscure spatial relationships. **(b)** Geometry-guided pruning may overcompress complementary views. **(c)** WaLKPruner refines retained evidence through scene abstraction and viewpoint tracking.

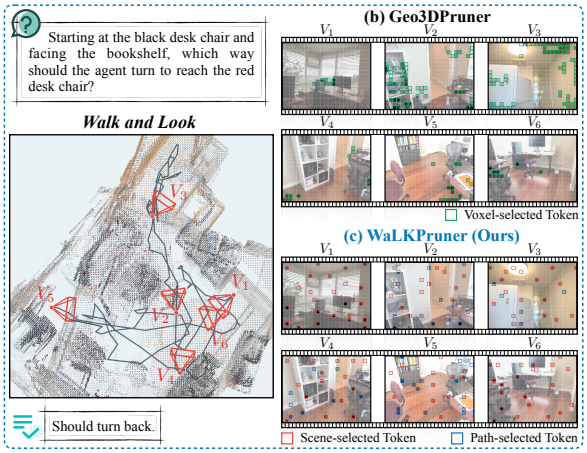


Figure 2: **Token visualization** for Geo3DPruner (Li et al., 2026a) (top) and WaLKPruner (bottom) at the same visual-token retention ratio. WaLKPruner retains more widely distributed evidence across viewpoints, including observations of the chairs, bookshelf, and their layout that support the turning decision.

- **Scene–trajectory pruning.** We introduce WaLKPruner, a training-free pruning framework for spatial understanding. It jointly models scene coverage and ordered viewpoint changes to retain representative content and complementary observations under a fixed token budget.
- **Path-guided refinement.** We formulate path-guided refinement that combines scene and path coverage to propose complementary observations from different stages of the camera trajectory. Constrained selective update incorporates them while limiting changes to the initial scene-selected set and bounding scene coverage loss.
- **Performance and analyses.** Experiments on VSI-Bench, ScanQA, and SQA3D show strong performance across token budgets. Token-allocation studies, visualizations, and ablations support using both scene content and observation order for evidence selection. Efficiency analyses on VSI-Bench show higher performance at lower computational cost than Geo3DPruner.

## 2 Related Work

**Spatial Understanding with MLLMs.** Recent multimodal large language models (MLLMs) have improved spatial understanding by incorporating geometric cues and spatial representations. Spatial-MLLM (Wu et al., 2025) and related approaches enhance scene understanding by aligning visual features with implicit 3D structures from geometry foundation models such as VGGT (Wang et al., 2025b). Other methods introduce explicit spatial representations, including 3D positional encoding or reconstruction objectives (Chen et al., 2024b; Ray et al., 2024; Wang et al., 2025a; Yang et al., 2026; Huang et al., 2025). Recent works further explore reasoning-based spatial understanding through chain-of-thought or reinforcement learning (Zheng et al., 2025; Ouyang et al., 2025; Chen et al., 2026; Fan et al., 2026). AffIn-Space (Lu et al., 2026) investigates spatial reasoning from the perspective of interaction-aware representations. These methods improve spatial perception by enhancing geometric or semantic representations, but generally focus on improving the representation capability of MLLMs rather than reducing redundant visual observations.

**Visual Token Pruning.** Visual token pruning reduces the computational cost of MLLMs by removing redundant visual tokens before language model inference. Existing approaches mainly rely on token importance, feature redundancy, or temporal consistency. FastV (Chen et al., 2024a), SparseVLM (Zhang et al., 2024), VisionZip (Yang et al., 2025b), and ZOO-Prune (Kim et al., 2026) select tokens based on attention or feature-based criteria, while FastVID (Shen et al., 2026) and FrameFusion (Fu et al., 2025) further exploit redundancy in video inputs. Geometry-aware methods such as Geo3DPruner (Li et al., 2026a) incorporate reconstructed spatial structures to maintain geometric diversity. However, feature-based pruning may remove spatially informative details, while geometry-based selection introduces reconstruction overhead and may overlook complementary observations from different viewpoints. In contrast, we consider visual tokens as observations along a camera trajectory and preserve evidence by jointly modeling scene content and viewpoint changes.

**Cognitively Inspired Visual Evidence Selection.** Human visual search combines rapid global perception with selective processing of informative regions under limited attention capacity (Wolfe et al., 2011). The nonselective pathway extracts coarse scene information in parallel, while the selective pathway focuses on task-relevant evidence guided by scene structure and attentional cues (Li et al., 2026b). Work on cognitive visual search, such as CVSearch (Li et al., 2026c), applies this principle to multimodal perception by combining global assessment with selective exploration. Inspired by this mechanism, we formulate token pruning as evidence selection rather than redundancy removal. Our framework constructs a compact scene abstraction and tracks viewpoint changes along the observation trajectory to identify complementary evidence. This cognitively inspired strategy enables efficient token allocation while preserving key spatial cues for scene understanding.

## 3 Method

### 3.1 Overview

**Overall Architecture.** As shown in Figure 3, `WaLKPruner` prunes visual tokens in a pretrained multimodal large language model (MLLM) to support spatial understanding under a fixed token budget. The visual input $\boldsymbol{x}_{\text{vis}} = (\boldsymbol{I}^{(1)}, \ldots, \boldsymbol{I}^{(T)})$ contains $T$ ordered video frames with a text question $\boldsymbol{x}_{\text{txt}}$. The MLLM contains a 2D visual encoder, a 3D geometry encoder, a tokenizer, a connector, and decoder layers. The connector fuses visual and geometry features into 3D-aware tokens $\boldsymbol{Z} = (\boldsymbol{z}_a)_{a=1}^{N}$, where $\boldsymbol{z}_a \in \mathbb{R}^d$ and $N$ is the token count before pruning. We directly reuse camera poses from the geometry encoder to track viewpoint changes, without constructing an additional voxelized scene representation for pruning. Given budget $B$, `WaLKPruner` retains $B$ fused tokens in original order before the MLLM decoder layers. The retained visual tokens and unchanged text tokens are then processed by the decoder layers to generate the spatial answer.

**Scene Abstraction and Viewpoint Tracking.** Our cognitive inspiration is to retain representative scene content while tracking changes in viewpoint. Scene abstraction selects tokens based on feature similarity, while viewpoint tracking accumulates camera translation and rotation along the observation sequence. Within viewpoint tracking, selective update prioritizes tokens from viewpoints adjacent to substantial pose changes. Joint scoring then combines scene and path coverage to

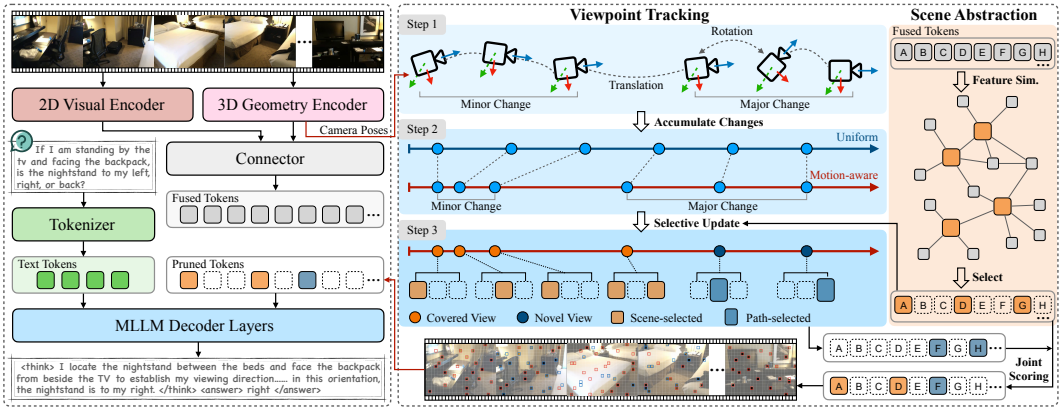


Figure 3: **Overview of `WaLKPruner`.** Our model combines scene abstraction and viewpoint tracking for training-free pruning. Scene abstraction selects representative content, while viewpoint tracking uses camera motion to guide selection. Selective updating and joint scoring refine the selection within a fixed budget for spatial question answering with the pretrained MLLM.

identify candidates and refines the initial selection according to these priorities. The selected tokens jointly represent scene content and different stages of the camera trajectory.

### 3.2 Scene Abstraction

**Feature Similarity.** Repeated observations can produce similar visual tokens across frames. Scene abstraction summarizes this content by retaining tokens that represent the full sequence. For token indices $a$ and $b$, we measure feature similarity using the nonnegative part of cosine similarity:

$$K^S_{ab} = \max\left(0, \frac{\boldsymbol{z}_a^\top \boldsymbol{z}_b}{\|\boldsymbol{z}_a\|_2 \|\boldsymbol{z}_b\|_2}\right). \tag{1}$$

For a selected index set $\mathcal{C} \subseteq \{1, \ldots, N\}$, we define scene coverage as

$$F_S(\mathcal{C}) = \frac{1}{N} \sum_{a=1}^{N} \max_{b \in \mathcal{C}} K^S_{ab}. \tag{2}$$

Each input token is represented by its most similar selected token, and the average similarity measures coverage of repeated visual content over the full sequence.

**Scene Selection.** We construct the initial scene-selected set $\mathcal{C}_S$ using greedy facility-location selection. Starting from an empty set with zero coverage, we repeatedly add the candidate with the largest marginal increase in $F_S$ among a sampled candidate pool until $|\mathcal{C}_S| = B$. Stochastic-greedy candidate sampling reduces selection cost, and a fixed random seed makes the procedure reproducible. The resulting set provides a compact representation of scene content and serves as the starting point for selective update guided by the ordered trajectory, preserving scene coverage.

### 3.3 Viewpoint Tracking

**Camera Translation and Rotation.** Feature similarity does not explicitly distinguish observations from different stages of a camera trajectory. Viewpoint tracking describes these stages using camera position, orientation, and temporal order. We consider $V$ ordered viewpoints after temporal merging and denote the source viewpoint of token $a$ by $v(a) \in \{1, \ldots, V\}$. The camera extrinsics of viewpoint $i$ yield its center $c_i$ and unit forward and right axes $f_i$ and $r_i$. We measure pose changes by combining normalized translation with rotation computed from the forward and right camera axes:

$$d_{ij} = \alpha \frac{\|c_i - c_j\|_2}{\max_{p,\ell} \|c_p - c_\ell\|_2} + (1 - \alpha) \frac{\arccos\langle f_i, f_j \rangle + \arccos\langle r_i, r_j \rangle}{2\pi}, \tag{3}$$

where $\alpha \in [0, 1]$ balances translation and rotation. Translation is normalized by the maximum distance between camera centers, while the two camera axes jointly describe orientation changes.

**Accumulate Changes.** Uniform temporal spacing does not reflect the amount of camera motion between observations. We therefore accumulate pose changes along the original temporal order to obtain the motion-aware coordinate illustrated in Figure 3. Let $e_i = d_{i,i+1}$ denote the change between adjacent viewpoints. With $s_1 = 0$, the normalized path position of viewpoint $i$ is $s_i = \sum_{k=1}^{i-1} e_k \,/\, \sum_{k=1}^{V-1} e_k$, where $i = 2, \ldots, V$. Small accumulated changes place observations close together, while larger changes separate them along the path. This coordinate preserves traversal order, including when the camera returns to a similar pose after substantial motion. It describes the sampled viewpoints without resampling the input video. We measure path coverage based on the distance from each viewpoint to the nearest viewpoint represented by the selected tokens along the accumulated path:

$$F_P(\mathcal{C}) = \frac{1}{V} \sum_{i=1}^{V} \max_{b \in \mathcal{C}} \exp\left( -\frac{(s_i - s_{v(b)})^2}{2\sigma^2} \right), \tag{4}$$

where $\sigma$ is determined from the median positive gap between adjacent path positions, without independent tuning. A viewpoint contributes a higher coverage score when a selected token originates from a nearby path position within the ordered trajectory.

**Selective Update.** Scene abstraction selects tokens that represent scene content. Selective updating uses camera motion to determine which viewpoints should receive priority when supplementing this initial selection. We define viewpoint importance as

$$q_i = \frac{e_{i-1} + e_i}{2 \sum_{k=1}^{V-1} e_k}, \quad i = 1, \ldots, V, \tag{5}$$

where $e_0 = e_V = 0$. This quantity assigns greater importance to viewpoints adjacent to larger translation or rotation changes. Each token $a$ inherits the importance $q_{v(a)}$ of its source viewpoint. For tokens outside the initial scene-selected set $\mathcal{C}_S$, this value determines their priority for insertion. Joint scoring uses this priority to process candidate additions and determine the final selection.

**Joint Scoring.** We identify candidate additions by considering scene and path coverage together. Starting from an empty set, we construct a size-$B$ candidate set $\mathcal{C}_J$ using greedy selection with marginal gains in

$$(1 - \lambda) F_S(\mathcal{C}) + \lambda F_P(\mathcal{C}), \tag{6}$$

where $\lambda \in [0, 1]$ controls the contribution of path coverage. At each step, the selected candidate increases the joint score the most among those evaluated. The set $\mathcal{C}_J$ provides candidates for refining $\mathcal{C}_S$, and only tokens in $\mathcal{C}_J \setminus \mathcal{C}_S$ are considered for insertion. We initialize the retained set with $\mathcal{C}_S$ and process these candidates in descending order of viewpoint importance $q_{v(a)}$. For each candidate, we find the currently retained token whose removal causes the smallest decrease in scene coverage. We then consider replacing that token with the candidate and accept the replacement only if it satisfies the following two constraints. Each proposed replacement is accepted only if the updated set contains at most $\lfloor \rho B \rfloor$ tokens from outside $\mathcal{C}_S$ and maintains scene coverage of at least $F_S(\mathcal{C}_S) - \epsilon$. Here, $\rho \in [0, 1]$ limits the updated fraction, and $\epsilon \geq 0$ bounds the coverage reduction relative to the fixed initial set $\mathcal{C}_S$. Each accepted replacement exchanges one token for another, so the final set $\mathcal{C}^\star$ contains exactly $B$ tokens while allowing the selection to change across viewpoints.

### 3.4 Inference Procedure

Given a retention ratio $r$, `WaLKPruner` retains the corresponding number of visual tokens using fused features and camera poses, independently of the text question. The same selector parameters are used across retention ratios within each dataset. We preserve the original order of the selected embeddings and remove unselected visual tokens and their attention-mask entries before the first MLLM decoder layer. The MLLM decoder layers process the retained visual tokens together with the unchanged text tokens to generate the spatial answer. All pretrained parameters remain fixed, and `WaLKPruner` requires no additional training or gradient computation.

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

**Datasets and Metrics.** We evaluate `WaLKPruner` on VSI-Bench (Yang et al., 2025a), ScanQA (Azuma et al., 2022), and SQA3D (Ma et al., 2022). Following the official evaluation protocols, we use the complete evaluation sets without question subsampling. VSI-Bench evaluates video-based spatial understanding across multiple tasks and reports the aggregated VSI Average score, which combines task-specific accuracy metrics. For ScanQA and SQA3D, we follow the standard evaluation scripts and report CIDEr and exact match (EM).

**Baselines.** We compare `WaLKPruner` with representative visual-token pruning approaches, including FastV (Chen et al., 2024a), ZOO-Prune (Kim et al., 2026), VisionZip (Yang et al., 2025b), SparseVLM (Zhang et al., 2024), FastVID (Shen et al., 2026), and FrameFusion (Fu et al., 2025). We further compare with the geometry-guided pruning method Geo3DPruner (Li et al., 2026a). For a fair comparison, we reproduce all baselines on the same Spatial-MLLM backbone and evaluate all methods at the same visual-token retention ratios.

**Implementation Details.** All experiments are conducted on an NVIDIA A100-PCIE-40GB GPU. `WaLKPruner` is built upon the Spatial-MLLM framework and uses the Spatial-MLLM-v1.1-Instruct-135K 3B checkpoint (Wu et al., 2025), with Qwen2.5-VL as the visual backbone and VGGT-1B (Wang et al., 2025b) as the geometry encoder. We sample 16 ordered frames for each video input and perform token selection on the fused 2D–3D visual tokens before the first MLLM decoder layer. The pruning procedure is question-independent and training-free, with all pretrained parameters frozen. For Scene Abstraction, we use stochastic-greedy facility-location selection, where candidates are selected according to their marginal increase in scene coverage. Joint Scoring follows the greedy strategy using the combined scene–path objective to construct candidates for Selective Updating. The path-kernel bandwidth $\sigma$ is determined from the median positive gap between adjacent path positions without tuning. We evaluate retention ratios of $r \in \{50\%, 25\%, 12.5\%, 6.25\%\}$. The default selection parameters are $\alpha = 0.60$ for pose distance, $\lambda = 0.02$ for joint scoring, $\rho = 0.05$ for selective updating, and $\epsilon = 0.002$ for the scene coverage constraint.

## 4.2 OVERALL PERFORMANCE

**Results on VSI-Bench.** As shown in Table 1, `WaLKPruner` achieves the best performance among all compared pruning methods across every token budget. With only 25% of the visual tokens, it reaches a VSI Average of 50.25, closely matching the unpruned Spatial-MLLM baseline of 50.39. Halving the retained budget from 50% to 25% reduces the score by only 0.04 points, suggesting substantial redundancy in the larger token set. The margin over Geo3DPruner grows from 0.86 points at 50% retention to 3.01 points at 6.25%, where `WaLKPruner` scores 46.70. This trend is consistent with path-guided refinement retaining complementary observations across viewpoints alongside representative scene content, supporting spatial understanding as the token budget becomes more constrained.

Table 1: **Evaluation on VSI-Bench.** VSI Average is reported across four visual-token retention ratios. `WaLKPruner` consistently outperforms existing pruning methods across all token budgets.

| Method | Year | 50% | 25% | 12.5% | 6.25% |
|---|---|---|---|---|---|
| FastV | 2024a | 48.24 | 46.22 | 44.49 | 43.21 |
| SparseVLM | 2024 | 45.62 | 43.70 | 42.27 | 40.89 |
| VisionZip | 2025b | 45.32 | 44.24 | 42.39 | 42.63 |
| FrameFusion | 2025 | 43.56 | 41.84 | 42.29 | 41.95 |
| ZOO-Prune | 2026 | 47.00 | 45.83 | 44.58 | 43.61 |
| FastVID | 2026 | 46.45 | 44.65 | 43.84 | 42.77 |
| Geo3DPruner | 2026a | 49.43 | 48.07 | 45.62 | 43.69 |
| **WaLKPruner** | 2026 | 50.29 | 50.25 | 48.39 | 46.70 |

**Results on ScanQA.** As shown in Table 2, `WaLKPruner` achieves the highest CIDEr at 50%, 25%, and 12.5% retention. At 25%, it leads both metrics with 86.63 CIDEr and 24.56 EM, exceeding Geo3DPruner by 3.52 CIDEr points. At 12.5%, its CIDEr exceeds FastVID by 1.50 points, while at 6.25% it ranks second on both metrics despite severe compression. Scores on both metrics exceed those of Geo3DPruner at every budget, with larger margins at 25% retention and below than at 50%. These results are consistent with scene abstraction preserving representative content and viewpoint tracking guiding the inclusion of complementary observations from underrepresented viewpoints, retaining useful spatial context for question answering under substantial token reduction.

Table 2: **Evaluation on ScanQA.** `WaLKPruner` leads CIDEr at most retention ratios and remains competitive on EM. These results are consistent with preserving representative scene content and complementary observations under limited token budgets.

| Method | 50% | | 25% | | 12.5% | | 6.25% | |
|---|---|---|---|---|---|---|---|---|
| | CIDEr | EM | CIDEr | EM | CIDEr | EM | CIDEr | EM |
| FastV (Chen et al., 2024a) | 79.37 | 22.41 | 74.87 | 21.24 | 72.16 | 20.71 | 70.98 | 20.28 |
| SparseVLM (Zhang et al., 2024) | 82.04 | 24.73 | 78.96 | 23.49 | 75.04 | 22.40 | 72.93 | 21.86 |
| VisionZip (Yang et al., 2025b) | 81.80 | 24.41 | 78.77 | 23.17 | 77.56 | 22.59 | 76.19 | 22.33 |
| FastVID (Shen et al., 2026) | 83.42 | 24.83 | 81.63 | 24.36 | 81.28 | 23.74 | 79.30 | 23.23 |
| Geo3DPruner (Li et al., 2026a) | 88.11 | 24.49 | 83.11 | 23.34 | 77.17 | 21.69 | 75.23 | 20.94 |
| **`WaLKPruner`** (Ours) | 88.13 | 24.58 | 86.63 | 24.56 | 82.78 | 23.44 | 78.58 | 22.40 |

Table 3: **Evaluation on SQA3D.** `WaLKPruner` leads both metrics at the lowest retention ratio and shows smaller declines than Geo3DPruner as compression increases. This trend is consistent with path-guided refinement preserving complementary spatial evidence.

| Method | 50% | | 25% | | 12.5% | | 6.25% | |
|---|---|---|---|---|---|---|---|---|
| | CIDEr | EM | CIDEr | EM | CIDEr | EM | CIDEr | EM |
| FastV (Chen et al., 2024a) | 137.05 | 51.73 | 133.08 | 50.32 | 129.58 | 48.94 | 127.53 | 48.05 |
| SparseVLM (Zhang et al., 2024) | 134.74 | 51.15 | 131.94 | 50.05 | 132.81 | 50.26 | 129.24 | 48.79 |
| VisionZip (Yang et al., 2025b) | 138.49 | 52.68 | 137.65 | 52.35 | 136.25 | 51.73 | 136.09 | 51.76 |
| FrameFusion (Fu et al., 2025) | 137.75 | 52.28 | 137.88 | 52.24 | 137.98 | 52.53 | 135.94 | 50.92 |
| FastVID (Shen et al., 2026) | 136.19 | 51.48 | 135.72 | 51.55 | 135.78 | 51.40 | 135.44 | 51.27 |
| Geo3DPruner (Li et al., 2026a) | 142.64 | 53.88 | 139.67 | 52.96 | 135.42 | 51.30 | 134.21 | 50.81 |
| **`WaLKPruner`** (Ours) | 141.49 | 53.42 | 139.94 | 52.74 | 138.23 | 52.19 | 137.44 | 51.93 |

**Results on SQA3D.** Table 3 shows that `WaLKPruner` becomes increasingly competitive as the token budget is reduced. Although Geo3DPruner leads at 50% retention, `WaLKPruner` achieves the highest CIDEr at the other three budgets and the best EM at 6.25%. At 6.25%, it obtains 137.44 CIDEr and 51.93 EM, exceeding the strongest baseline, VisionZip, by 1.35 and 0.17 points, respectively. Its CIDEr drops by 4.05 points from 50% to 6.25%, compared with 8.43 for Geo3DPruner. EM shows the same pattern, dropping by 1.49 points for `WaLKPruner` and 3.07 for Geo3DPruner over this range. The smaller declines in both metrics are consistent with jointly preserving representative scene content and complementary observations, helping maintain spatial references as the token budget becomes more constrained and reversing the ranking under stronger compression.

### 4.3 Quantitative and Qualitative Analyses

**Efficiency–efficacy trade-off.** In Figure 4, `WaLKPruner` consistently occupies the upper-left region relative to Geo3DPruner, achieving stronger spatial reasoning with lower computational cost. At 25% retention, `WaLKPruner` achieves a VSI Average of 50.25 with 164.14 TFLOPs, compared with 48.07 and 170.95 TFLOPs for Geo3DPruner. As the token budget increases, performance improves rapidly and then saturates, indicating that retaining additional tokens brings diminishing benefits once representative scene content and complementary viewpoints have been preserved. This trend aligns with our design: scene abstraction removes redundant observations, while viewpoint tracking preserves complementary evidence along the camera path, allowing the limited token budget to focus on spatially informative content.

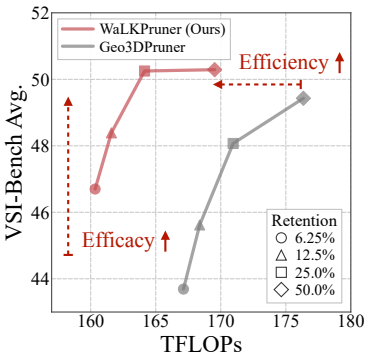


Figure 4: **Efficiency–efficacy trade-off** on VSI-Bench.

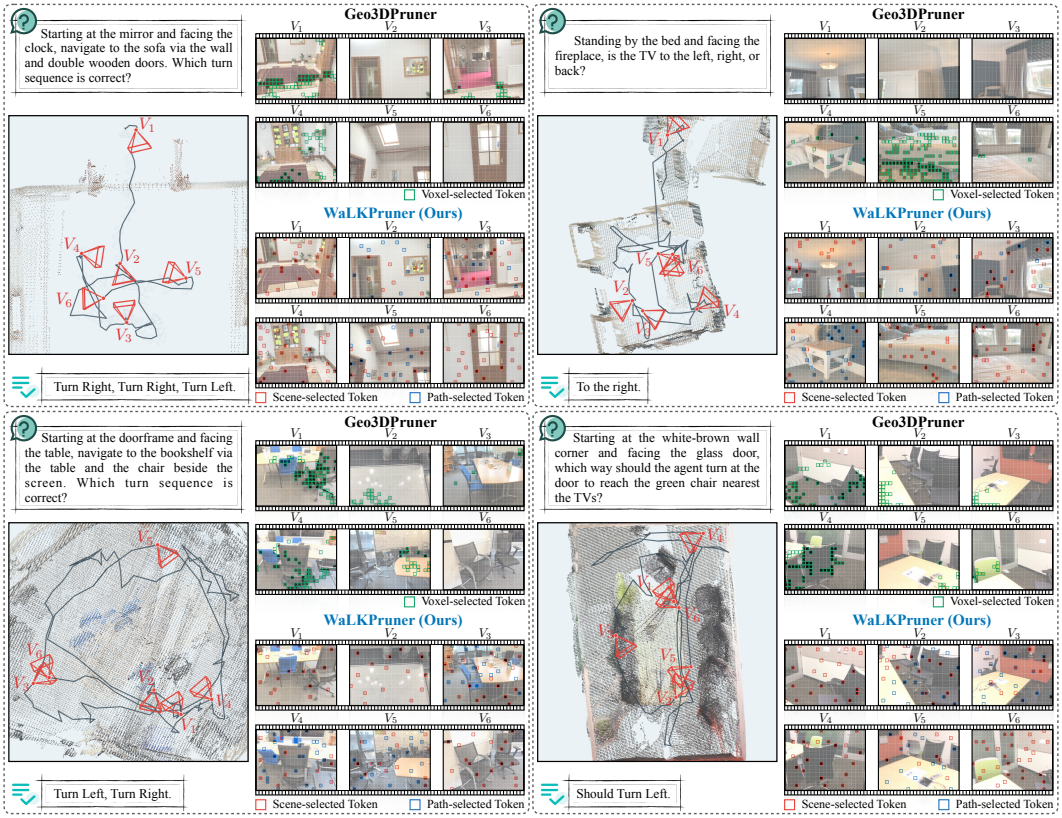


Figure 6: **Qualitative comparison** across four cases of turn sequence prediction and relative direction reasoning. Each case shows the scene and observation trajectory alongside tokens retained by Geo3DPruner (Li et al., 2026a) (top) and `WaLKPruner` (bottom) at the same visual-token retention ratio of 6.25%. `WaLKPruner` combines scene-selected (red) and path-selected (blue) tokens to retain more widely distributed evidence across viewpoints, supporting the spatial judgments.

**View-wise token allocation.** At 25% visual-token retention, Figure 5 examines how the fixed token budget is distributed across individual views. Both methods have the same mean budget of 97.8 tokens per view, yet their allocation patterns differ substantially. Geo3DPruner spans a much wider observed range of 0–319 tokens per view, whereas `WaLKPruner` narrows this range to 33–206 and concentrates more views around the shared mean. Rather than enforcing uniform allocation, viewpoint tracking preserves support for observations along the camera path, while scene abstraction avoids spending excessive tokens on redundant content. This allocation reduces extreme per-view budgets and maintains broader viewpoint coverage without increasing the overall token budget.

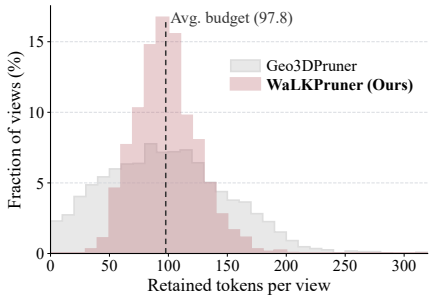


Figure 5: **View-wise token allocation at 25% retention.** Distributions of retained tokens per view for Geo3DPruner and `WaLKPruner`. The dashed line marks their shared average budget of 97.8 tokens per view.

**Case study.** Figure 6 presents four cases with scenes visualized using VGGT-SLAM (Maggio et al., 2026). Geo3DPruner often concentrates tokens in particular viewpoints or regions, while `WaLKPruner` retains scene content together with complementary observations along the path. In the upper cases, these observations provide spatial references for relating the mirror and clock to the sofa, and the TV to the bed and fireplace. The lower-left case illustrates that nearby camera positions can yield different views of the table and chairs, with path-selected tokens in $V_3$ complementing scene-selected tokens in $V_6$ along the looping path. In the lower-right case, path-selected tokens in $V_2$, $V_3$, and $V_5$ retain views of the table, chairs, and screens, complementing scene-selected observations of the wall

corner, glass door, and green chair. These examples illustrate how scene abstraction and viewpoint tracking jointly preserve spatial references across changes in viewing direction.

### 4.4 Ablation Study

**Complementarity of scene and trajectory cues.** Figure 7 examines scene representativeness and ordered viewpoint trajectories for spatial token pruning across retention ratios. Removing either scene abstraction or viewpoint tracking consistently degrades the VSI Average, supporting their complementary roles. Without viewpoint tracking, feature-based scene coverage alone cannot fully retain observations revealed along the trajectory. Removing scene abstraction lowers performance, showing that trajectory cues cannot replace a compact, representative scene summary. Replacing the ordered trajectory with poseset, which retains the same poses but discards observation order, yields smaller yet consistent degradation. Because both variants use identical camera poses, this gap cannot be attributed to pose diversity alone. Modeling ordered viewpoint transitions captures how complementary evidence emerges along the trajectory, enabling constrained refinement of scene-selected tokens.

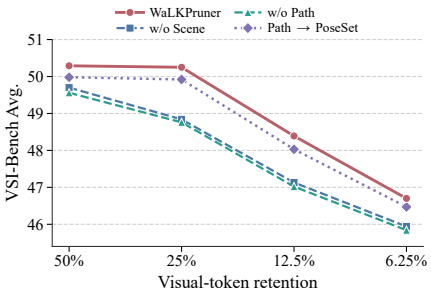


Figure 7: **Scene and trajectory ablations on VSI-Bench.** Removing either scene abstraction or viewpoint tracking degrades performance, while poseset reveals a consistent benefit from preserving trajectory order across budgets.

**Parameter sensitivity.** Figure 8 evaluates the translation–rotation balance $\alpha$, path weight $\lambda$, and selective-update ratio $\rho$ at 25% retention. Across the tested ranges, the VSI Average varies by approximately 0.6 points or less, indicating limited sensitivity to these parameters. For $\alpha$, performance remains stable as the balance between translation and rotation changes, suggesting that viewpoint tracking does not require precise tuning of their relative contributions. For $\lambda$, performance peaks at a modest path weight and decreases slightly toward either end of the tested range. This pattern is consistent with trajectory cues supplying complementary observations along the camera path to supplement the representative content provided by scene abstraction. For $\rho$, the best result occurs with a small update allowance, while larger allowances produce no further gains. This supports conservative refinement under the scene coverage constraint, which limits disruption to the representative initial selection. The defaults $\alpha = 0.60$, $\lambda = 0.02$, and $\rho = 0.05$ achieve the highest VSI Average in their respective sweeps.

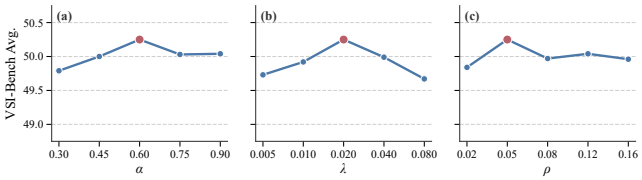


Figure 8: **Parameter sensitivity on VSI-Bench at 25% token retention.** Each panel varies one parameter while keeping the others fixed. Red markers denote the best-performing defaults across parameters.

## 5 Conclusions

We presented **`WaLKPruner`**, a training-free framework for cognitively inspired visual token pruning in spatial understanding. Our approach connects a compact scene overview with complementary observations from changing viewpoints, using scene abstraction and viewpoint tracking to guide selection along an ordered observation path. Constrained selective updates refine this selection under a fixed token budget while limiting changes to the initial scene representation. Experiments on three benchmarks demonstrate strong performance across token budgets, while analyses on VSI-Bench show a favorable balance between computational cost and spatial understanding. Token allocation studies reveal a more balanced distribution across viewpoints, while case studies illustrate the spatial references provided by complementary observations. Ablations further support the joint use of scene content and path information, including the benefit of retaining observation order when the compared variants use the same camera poses. These findings support using ordered viewpoint changes to complement scene content when preserving key visual cues for spatial understanding.

## Reproducibility Statement

We point readers to the Method (Section 3), including the inference procedure (Section 3.4), and to the Experimental Setup (Section 4.1), which summarizes the model configuration, implementation settings, token budgets, and selection parameters. These sections contain the information needed to reproduce our results. We will release the source code and evaluation scripts upon acceptance.

## LLM Usage Statement.

Consistent with policies on LLM usage, we used LLMs to assist with experimental code organization, checking the consistency and fairness of experimental comparisons, and language polishing. All research ideas, experiments, results, and analyses were produced and verified by the authors, who take full responsibility.

## References

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19129–19139, 2022.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

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*, pp. 19–35. Springer, 2024a.

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding reasoning and planning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26428–26438, 2024b.

Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Xiang An, Yan Feng, Peng Pei, Xunliang Cai, et al. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2613–2624, 2026.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024c.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Peihao Wang, Huaizhi Qu, Shijie Zhou, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31054–31065, 2026.

Tianyu Fu, Tengxuan Liu, Qinghao Han, Guohao Dai, Shengen Yan, Huazhong Yang, Xuefei Ning, and Yu Wang. Framefusion: Combining similarity and importance for video token reduction on large vision language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22654–22663. IEEE, 2025.

Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. 3DRS: MLLMs need 3d-aware representation supervision for scene understanding. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Youngeun Kim, Youjia Zhang, Huiling Liu, Aecheon Jung, Sunwoo Lee, and Sungeun Hong. Zooprune: training-free token pruning via zeroth-order gradient estimation in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39572–39582, 2026.

Han Li, Zehao Huang, Jiahui Fu, Naiyan Wang, and Si Liu. Geometry-guided 3d visual token pruning for video-language models. *arXiv preprint arXiv:2604.18260*, 2026a.

Kailing Li, Qi'ao Xu, Tianwen Qian, Yuqian Fu, Yang Jiao, and Xiaoling Wang. Clivis: Unleashing cognitive map through linguistic-visual synergy for embodied visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5134–5143, 2026b.

Liupeng Li, Haoqian Kang, Zhenyu Lu, Jinpeng Wang, Bin Chen, Ke Chen, and Yaowei Wang. Cvsearch: Empowering multimodal llms with cognitive visual search for high-resolution image perception. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026c.

Zhenyu Lu, Liupeng Li, Jinpeng Wang, Haoqian Kang, Manyuan Zhang, Yan Feng, Ke Chen, and Yaowei Wang. Affin-space: Learning affine-invariant representations for 3d spatial understanding with mllms. In *International Conference on Machine Learning (ICML)*, 2026.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022.

Dominic Maggio, Hyungtae Lim, and Luca Carlone. Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. *Advances in Neural Information Processing Systems*, 38:129839–129867, 2026.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, et al. Sat: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.

Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, et al. Fastvid: Dynamic density pruning for fast video large language models. *Advances in Neural Information Processing Systems*, 38:123553–123581, 2026.

Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang Zhang. Ross3d: Reconstructive visual instruction tuning with 3d-awareness. *arXiv preprint arXiv:2504.01901*, 2025a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025b.

Jeremy M Wolfe, Melissa L-H Võ, Karla K Evans, and Michelle R Greene. Visual search in scenes involves selective and nonselective pathways. *Trends in cognitive sciences*, 15(2):77–84, 2011.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025a.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19792–19802. IEEE, 2025b.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Towards spatial supersensing in video. In *The Fourteenth International Conference on Learning Representations*, 2026.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024.

Duo Zheng, Shijia Huang, Yanyang Li, and Liwei Wang. Learning from videos for 3d world: Enhancing MLLMs with 3d vision geometry priors. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.