BEYOND SEMANTIC MATCHING: LET MODALITIES TALK THROUGH THEIR POSITION–CONTENT DUAL MANIFOLDS FOR VLM TOKEN PRUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual token pruning is an effective way to reduce the inference cost of VLMs, yet identifying a compact and sufficient subset of visual tokens remains challenging, as token importance depends not only on cross-modal relevance but also on complex, heterogeneous **Where–What** relations in each modality. Existing methods either prematurely mix position and content information into a shared representation or model them in separate stages, potentially obscuring their distinct structures and intrinsic dependency-relations. To address this, we introduce **Position–Content Dual Manifold (PCDM)**, a unified multimodal, multi-attribute relational system for visual token pruning. PCDM factorizes each token into coupled position and content nodes, thus forming two complementary manifolds per modality: a position manifold preserving native structural proximity and a content manifold capturing semantic correlations, with one-to-one node correspondence linking them together. The vision and language PCDMs are further connected through semantic alignment, enabling information to propagate globally across modalities and attributes while preserving their distinct structures. This enables inference of higher-order task relevance and structured pairwise affinities among visual tokens. Token selection is then formulated as a budget-constrained quadratic optimization that rewards task relevance while suppressing semantic/spatial redundancy. Experiments across *five VLMs and ten benchmarks* demonstrate consistently superior efficiency–accuracy trade-offs over state-of-the-art methods. On LLaVA-NeXT-7B, PCDM retains 94.5% of full performance with only 160 visual tokens (5.6%).

1 INTRODUCTION

Vision–Language Models (VLMs) can incur substantial computational overhead when processing large numbers of visual tokens under high-resolution settings (Jeddi et al., 2025). As Transformer computation grows rapidly with sequence length (Vaswani et al., 2017), *visual token pruning* (Liu et al., 2023) has emerged as a direct and effective approach to improving inference efficiency.

Effective visual token pruning requires systematically evaluating token informativeness and relational structure, which in VLMs are shaped not only by cross-modal interactions between vision and language, but also by heterogeneous attributes – *position and content* – and their interactions within each modality (e.g., 2D spatial layout and visual semantics in images, or 1D sequential order and linguistic semantics in text). This calls for a unified treatment of higher-order token relations across both modalities and attributes. Under prevailing ViT-based architectures, however, such unified modeling faces two key challenges.

First, modeling intrinsic dependency between token position and content while preserving their distinct structures is challenging. Mainstream ViT encoders typically couple the two attributes at an early stage. Additive positional embeddings directly mix them, leading to noisy correlations between the two attributes (Ke et al., 2021). RoPE couples the two through position-dependent rotations, making their interaction implicit and difficult to interpret. In both cases, flattening 2D image grids into a 1D token sequence will also obscure fine-grained spatial relations.

Recent work has increasingly recognized the importance of visual-token positions in VLM token pruning. (Zou et al., 2026; Duan et al., 2026; Huang et al., 2026b) adopt predefined spatial par-

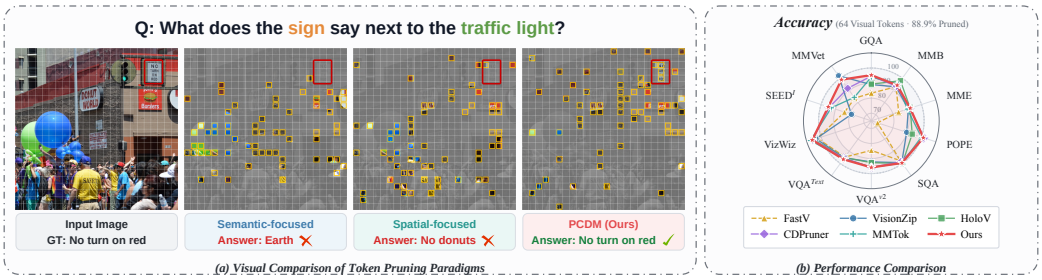


Figure 1: **Overview of PCDM.** (a) Preserved visual tokens by different methods. PCDM provides correct answer even under aggressive compression. (b) Comparative results on benchmark datasets with LLaVA-1.5-7B with 64 visual tokens (88.9% pruning). PCDM shows encouraging results.

titions with adaptive region-wise token budgets to avoid spatial crowding and improve coverage. (Sun et al., 2026; Lu et al., 2026) exploit positional priors, such as RoPE phase information, to introduce spatial awareness into token retention. However, relying on regional constraints or RoPE as indirect proxies may partially obscure native, fine-grained 2D spatial relationships. Moreover, these additional positional cues are typically introduced and modeled separately from content semantics, leaving the intrinsic interaction between *where* and *what* insufficiently captured.

Second, cross-modal composition further amplifies the complexity of heterogeneous token relations. The importance of a visual token depends not only on its spatial and content relations with other visual tokens, but also on its semantic alignment with the prompt, which is a higher-order cross-modal relational system rather than a collection of independent local cues. Existing methods typically capture such interactions at a relatively local level, without fully propagating and integrating relational signals over the entire global structure. Consequently, long-range cross-modal dependencies involving heterogeneous attributes may not be fully resolved.

Solving both challenges requires *preserving where–what information while still allowing them to freely interact with each other*. To this end, we propose **Position–Content Dual Manifold (PCDM)**, a structured relational modeling system for VLM token pruning. PCDM starts from a simple observation: every token carries two complementary attributes—*where* it is located and *what* it represents. The former reflects the native structural organization (2D images or 1D sequences), and the the latter is encoded by high-dimensional features (patch or word/subword embeddings).

Based on this observation, PCDM factorizes each token into a pair of interconnected nodes: a *position node* encoding its where information and a *content node* encoding its what information, as illustrated in Fig. 2. For each token, the two nodes are connected by a one-to-one correspondence edge. Across tokens, position nodes are connected according to their structural proximity, while content nodes connected according to their semantic relationships. This naturally organizes each modality into a **dual-manifold relational system** with three components:

- **(i) Position Manifold,** which captures the native structural proximity among tokens induced by the spatial, sequential, or temporal organization of the modality.
- **(ii) Content Manifold,** which captures semantic relationships among tokens induced by correlations in their high-dimensional content representations.
- **(iii) Position–Content Correspondence,** which couples the two manifolds through one-to-one link between the position and content nodes of each token, allowing the two relational structures to interact while remaining explicitly separated.

Here, *manifold* refers to the relational structure induced by node-level graphs within each attribute domain (position, or content), rather than a strict differential-geometric manifold.

PCDM provides a flexible relational treatment of **where–what** information: rather than forcing positional and semantic attributes—which differ substantially in form and dimensionality – into a shared feature space and mix them prematurely, PCDM preserves them as distinct relational structures while enabling flexible interaction through explicit position–content correspondence. This avoids information collapse from additive feature fusion and allows spatial proximity and semantic similarity to reinforce or counterbalance each other at the relational level.

PCDM can naturally extend to multiple modalities. As illustrated in Figure 2, once vision and language are represented by their respective position–content dual manifolds, their content submanifolds can be further connected through cross-modal semantic edges, while the modality-specific positional structures remain unaffected within their original modalities. This yields a **global heterogeneous relational system**, in which token relations are jointly shaped by position, content, vision, and language, enabling global information flow to capture higher-order task relevance and structured pairwise token affinities. These signals are then jointly incorporated into a cardinality-constrained quadratic program to select a compact, representative, and non-redundant subset of visual tokens.

We next introduce related work, the technical details of PCDM, and report comprehensive evaluations results against SOTA training-free VLM token pruning methods.

2 Related Work

In this paper we focus on **training-free methods for token pruning.** For ease of discussion, we categorize existing methods into two categories by the information sources used for pruning.

Semantic-importance-guided pruning methods primarily address *what visual information is relevant* (Xing et al., 2024; Alvar et al., 2025a; Fang et al., 2026a; Li et al., 2026). Early methods such as FastV (Chen et al., 2024) estimate visual-token importance from LLM self-attention over the jointly encoded visual–text sequence. SparseVLM (Zhang et al., 2024b) propose to retain only visually relevant text tokens and use the text-to-vision attention to score visual tokens. FasterVLM (Zhang et al., 2024a) further simplifies the rating by replacing multiple text-conditioned raters with the vision encoder's `[CLS]` token, providing a compact visual-saliency estimate. Beyond task relevance, subsequent methods take into account extra constraints on visual tokens. VisionZip (Yang et al., 2025) suppresses redundant information by merging feature-similar vision tokens, while CD-Pruner (Zhang et al., 2026) and MMTok (Dong et al., 2025) promote diversity and coverage through Determinantal Point Process optimization and greedy maximum-coverage selection, respectively.

Spatial-position-aware pruning methods explicitly address *where the visual evidence should be preserved* (Chien et al., 2025; Huang et al., 2026a; Lu et al., 2026). HoloV (Zou et al., 2026) organizes visual tokens into local crops and adaptively allocates budgets to preserve spatial granularity and scene coverage; GridPrune (Duan et al., 2026) partitions the 2D grids into fixed regions and performs text-conditioned token selection locally with allocated budget. Nüwa (Huang et al., 2026b) maintains a global spatial reference through regional separation, representative-anchor alignment, and spatially weighted neighborhood aggregation. Beyond spatial partitioning, IVC-Prune (Sun et al., 2026) exploits the phase structure of RoPE to recover implicit visual-coordinate anchors.

Taken together, existing methods treat position and content either within a mixed representation or as separate cues introduced at different stages. In the former, spatial and semantic structures are mixed irreversibly, making it difficult to attribute their respective contributions. So semantic-driven selection may favor contextually dominant regions while overlooking spatially complementary evidence (Darcet et al., 2024; Kang et al., 2025). In the latter, the intricate interaction between position and content is only weakly modeled, leaving where-what dependency insufficiently captured, especially when task relevance need to propagate across both attributes and modalities.

PCDM is aimed to fill this gap. Its key distinction from existing methods is not simply to integrate position and content information, but to preserve them as **distinct while explicitly interacting** relational structures, enabling joint reasoning over *Where-What* dependencies in and across modalities. This gives a unified relational basis for estimating task relevance and redundancy for token pruning.

3 Position-Content Dual Manifolds for VLM Token Selection

We formulate visual token pruning as a *query-conditioned representative selection* problem over a global vision–language relational system constructed using PCDM. It consists of three stages:

(1) Global coupled PCDM construction. Construct a position–content dual manifold for each modality (vision and language), and connect the two PCDMs to form a global relational system.

(2) Query-guided relevance propagation. Propagate text-prompt signals throughout the coupled PCDM manifolds to evaluate system-wide, higher-order relevance scores for visual tokens.

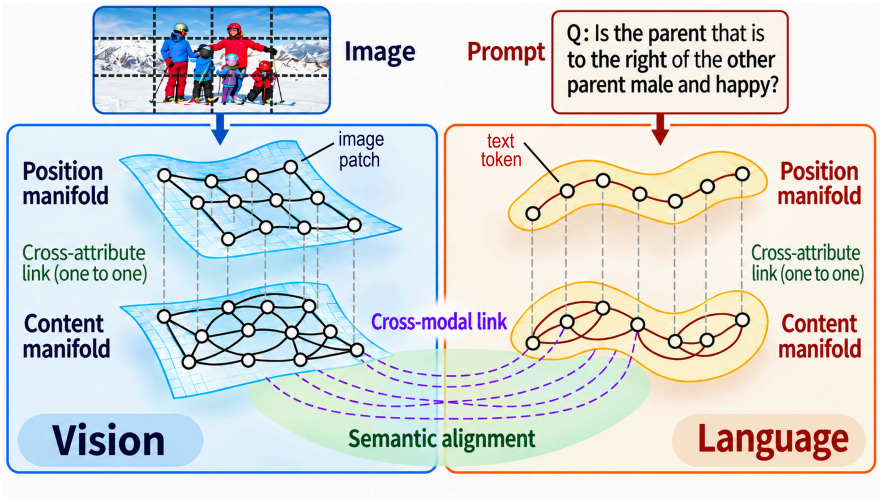


Figure 2: Overview of PCDM's position–content dual manifolds across vision & language.

(3) Relevance–redundancy token selection. Solve a quadratic program that maximizes the query relevance of the selected tokens while penalizing their pairwise semantic and spatial redundancy.

3.1 Global Position–Content Dual Manifold Construction

Given an image–query pair with n **visual tokens** and m **textual tokens**, we will build a Global PCDM relational system that spans both modalities, as well as their position-content attributes.

This global relational system can be cast as heterogeneous graph $\mathcal{G}$ with four types of nodes:

Text modality: m content-nodes $U_c = \{u_1^c, \ldots, u_m^c\}$, m position-nodes $U_p = \{u_1^p, \ldots, u_m^p\}$.

Vision modality: n content-nodes $V_c = \{v_1^c, \ldots, v_n^c\}$, n position-nodes $V_p = \{v_1^p, \ldots, v_n^p\}$.

The links between these $2(m+n)$ nodes, ordered as `text-content, text-position, vision-content, vision-position`, can be written as the following adjacency matrix

$$\mathbf{P} = \begin{bmatrix} \mathbf{W}^{\text{c-c}}_{m\times m} & \eta\mathbf{I}_{m\times m} & \mathbf{M}^{\text{c-c}}_{m\times n} & \mathbf{0} \\ \eta\mathbf{I}_{m\times m} & \mathbf{W}^{\text{p-p}}_{m\times m} & \mathbf{0} & \mathbf{0} \\ \left(\mathbf{M}^{\text{c-c}}_{m\times n}\right)^\top & \mathbf{0} & \mathbf{K}^{\text{c-c}}_{n\times n} & \eta\mathbf{I}_{n\times n} \\ \mathbf{0} & \mathbf{0} & \eta\mathbf{I}_{n\times n} & \mathbf{K}^{\text{p-p}}_{n\times n} \end{bmatrix}. \tag{1}$$

Here, $\mathbf{W}$ denotes within-language adjacency matrices: $\mathbf{W}^{\text{c-c}}$ captures pairwise semantic relationships among text tokens and $\mathbf{W}^{\text{p-p}}$ encodes their sequential neighborhood structure. Similarly, $\mathbf{K}$ denotes within-vision adjacency matrices: $\mathbf{K}^{\text{c-c}}$ captures pairwise semantic correlations among image patches and $\mathbf{K}^{\text{p-p}}$ encodes their native 2D spatial proximity using an 8-neighborhood (we use $\mathrm{Nbr}_8(j)$ to denote the 8-neighborhood of visual token j on the 2D grid). Finally, $\mathbf{M}^{\text{c-c}}_{m\times n}$ denotes the cross-modal adjacency matrix between the m textual tokens and n visual tokens.

Let $\bar{\boldsymbol{u}}_i$ and $\bar{\boldsymbol{v}}_j$ denote the (L_2-normalized) language and visual token embeddings, respectively. Let indices $i, i' \in \{1, \ldots, m\}$ refer to language tokens, and $j, j' \in \{1, \ldots, n\}$ refer to visual tokens. The non-empty blocks in eq. (2) are defined as

$$\begin{gathered} \mathbf{W}^{\text{c-c}}_{m\times m}(i,i') = \exp\big(\langle \bar{\boldsymbol{u}}_i, \bar{\boldsymbol{u}}_{i'}\rangle\big), \quad \mathbf{K}^{\text{c-c}}_{n\times n}(j,j') = \exp\big(\langle \bar{\boldsymbol{v}}_j, \bar{\boldsymbol{v}}_{j'}\rangle\big), \quad \mathbf{M}^{\text{c-c}}_{m\times n}(i,j) = \exp\big(\langle \bar{\boldsymbol{u}}_i, \bar{\boldsymbol{v}}_j\rangle\big), \\ \mathbf{W}^{\text{p-p}}_{m\times m}(i,i') = \eta\,\mathbb{I}_{\{i'\in[i-1,i+1]\}}, \qquad \mathbf{K}^{\text{p-p}}_{n\times n}(j,j') = \eta\,\mathbb{I}_{\{j'\in \mathrm{Nbr}_8(j)\}}. \end{gathered} \tag{2}$$

We set $\eta = e$ to match the numerical ranges of positional and semantic affinities. All content adjacency sub-matrices are sparsified at column-wise mean $+3$ standard deviations, followed by column normalization of $\mathbf{P}$ to obtain $\widetilde{\mathbf{P}}$. It provides a principled and interpretable decomposition of **Where–What** information, spanning both heterogeneous token relationships and different modals.

3.2 TEXT-GUIDED RELEVANCE DIFFUSION

To estimate the task relevance of visual tokens, we treat the language-content nodes as the source of task signals and diffuse them over the global PCDM system. Starting from the textual content manifold, the signal traverses the cross-modal links to the visual modality and propagates throughout the coupled manifolds. We initialize the distribution uniformly over m language-content nodes as

$$\boldsymbol{r}^{(0)} = \left[\boldsymbol{r}^{(0)}_{U_c},\ \mathbf{0},\ \mathbf{0},\ \mathbf{0}\right]^\top, \qquad r^{(0)}_{U_c,i} = \frac{1}{m}, \quad i = 1, \ldots, m. \tag{3}$$

Then we perform a random walk with restart over the global PCDM system. Given the column-normalized transition matrix $\widetilde{\mathbf{P}}$, the iterations are defined as $\boldsymbol{r}^{(t)} = (1-\alpha)\widetilde{\mathbf{P}}\boldsymbol{r}^{(t-1)} + \alpha\boldsymbol{r}^{(0)}$ where $\alpha \in (0,1)$ denotes the restart probability. The restart mechanism keeps the propagation anchored to the textual query while allowing the task signal to diffuse across the coupled relational system. Upon convergence, the steady-state distribution admits the closed-form

$$\boldsymbol{r}^\star = \alpha\left(\mathbf{I} - (1-\alpha)\widetilde{\mathbf{P}}\right)^{-1}\boldsymbol{r}^{(0)}. \tag{4}$$

Solving Eq. (4) requires inverting a $(2m+2n)\times(2m+2n)$ matrix. Although $\widetilde{\mathbf{P}}$ is sparse, the overhead can still accumulate when VLM needs to process large numbers of visual inputs. We therefore apply a low-rank CUR decomposition (Donello et al., 2023) to approximate $\widetilde{\mathbf{P}}$. It selects a small subset of the nodes with indices $\mathcal{L}$ from all the $2(m+n)$ nodes such that $|\mathcal{L}| << 2m+2n$, corresponding to a subset of columns (C), a subset of rows (R), and their intersections (U),

$$\mathbf{C} = \widetilde{\mathbf{P}}_{:,\mathcal{L}}, \qquad \mathbf{U} = \widetilde{\mathbf{P}}_{\mathcal{L},\mathcal{L}}, \qquad \mathbf{R} = \widetilde{\mathbf{P}}_{\mathcal{L},:}. \tag{5}$$

Here, $\mathbf{C} \in \mathbb{R}^{(2m+2n)\times\ell}$, $\mathbf{U} \in \mathbb{R}^{\ell\times\ell}$, and $\mathbf{R} \in \mathbb{R}^{\ell\times(2m+2n)}$. Then $\widetilde{\mathbf{P}}$ can be approximated as

$$\widetilde{\mathbf{P}} \approx \mathbf{C}\mathbf{U}^\dagger\mathbf{R}. \tag{6}$$

where $\mathbf{U}^\dagger$ is the pseudoinverse. Using Eq. (6), the r-star in Eq. (4) can be approximated by

$$\boldsymbol{r}^\star \approx \alpha\left[\mathbf{I}_{2m+2n} + (1-\alpha)\mathbf{C}\mathbf{H}^{-1}\mathbf{U}^\dagger\mathbf{R}\right]\boldsymbol{r}^{(0)}, \tag{7}$$

where $\mathbf{H}$ is the intermediate matrix computed through Woodbury formula,as

$$\mathbf{H} = \mathbf{I} - (1-\alpha)\mathbf{U}^\dagger\mathbf{R}\mathbf{C}. \tag{8}$$

Equation (7) replaces the inversion of a $(2m+2n)\times(2m+2n)$ matrix with that of the much smaller $\ell\times\ell$ matrix $\mathbf{H}$. The full derivation and landmark sampling scheme are provided in Appendix B.

Recall that $\boldsymbol{r}^\star$ provides relevance scores at the node level. Since each visual token is represented in PCDM by two corresponding nodes — a visual-content node and a visual-position node — we aggregate their relevance scores to obtain the token-level relevance, as

$$\boldsymbol{r}_V = a\,\boldsymbol{r}^\star[V_c] + (1-a)\,\boldsymbol{r}^\star[V_p]. \tag{9}$$

Here, $a \in (0,1)$ balances the contributions of content and positional evidence. The resulting $\boldsymbol{r}_V$ integrates text-conditioned semantic relevance with spatial structure and serves as the relevance term in the subsequent visual-token selection.

3.3 RELEVANCE–REDUNDANCY TOKEN PRUNING

Besides estimating token relevance, we also need pairwise relations to suppress redundant selections. To this end, we use the global PCDM system to derive *higher-order visual-token affinities* jointly capturing visual semantics, spatial structure, and language-conditioned information. Here, affinity between two visual tokens is characterized by the probability mass propagated from one to the other through the global, restarted random-walk diffusion on global PCDM system. Mathematically, all such propagations can be obtained from the closed-form diffusion matrix $\alpha(\mathbf{I} - (1-\alpha)\widetilde{\mathbf{P}})^{-1}$ - we can extract the content-content block and the position-position block within the diffusion matrix and combine them together as token-level affinity scores,

$$\mathbf{S}^{\text{global}} = (1-\alpha)\left[(1-b)\mathbf{C}[V_c,:]\mathbf{H}^{-1}\mathbf{U}^\dagger\mathbf{R}[:,V_c] + b\mathbf{C}[V_p,:]\mathbf{H}^{-1}\mathbf{U}^\dagger\mathbf{R}[:,V_p]\right]. \tag{10}$$

We reuse the CUR in Eq. (5) for efficient computation. Here, b balances content- and position-based affinities, and $\mathbf{S}^{\text{global}}$ is symmetrized with its transpose to obtain $\widetilde{\mathbf{S}}^{\text{global}}$ for the subsequent QP.

Since diffusion-based affinities emphasize higher-order relational dependencies and may smooth over direct local similarities, we retain the original visual affinity for redundancy estimation, by extracting the semantic similarity block ($\mathbf{K}^{\text{c-c}}_{n\times n}$) and spatial proximity block ($\mathbf{K}^{\text{p-p}}_{n\times n}$) directly from the global adjacency matrix Eq. (1), and mixing them with the same weighting factor b as in Eq. (9):

$$\mathbf{S}^{\text{ori}} = (1-b)\mathbf{K}^{\text{c-c}}_{n\times n} + b\mathbf{K}^{\text{p-p}}_{n\times n}. \tag{11}$$

The two redundancy matrices are integrated through a multiplicative gate

$$\mathbf{S} = \mathbf{S}^{\text{ori}} \odot \left(\mathbf{1}_{n\times n} + \gamma\widetilde{\mathbf{S}}^{\text{global}}\right), \tag{12}$$

where $\gamma \geq 0$ controls the strength of the global modulation - a token pair shows high redundancy in both $\mathbf{S}^{\text{ori}}$ and $\widetilde{\mathbf{S}}^{\text{global}}$, then the integration would induce a larger redundancy penalty.

Using task-relevance score $\boldsymbol{r}_V$ and redundancy matrix $\mathbf{S}$, visual token selection is formulated as a budget-constrained quadratic program (QP) (Kochenberger et al., 2014):

$$\max_{\boldsymbol{x}} \quad \boldsymbol{r}_V^\top \boldsymbol{x} - \lambda \boldsymbol{x}^\top \mathbf{S}\boldsymbol{x}, \qquad \text{s.t.} \quad \mathbf{1}_n^\top \boldsymbol{x} = K, \quad x_j \in \{0,1\},\ j = 1,\ldots,n. \tag{13}$$

Here, both $\boldsymbol{r}_V$ and $\mathbf{S}$ are normalized to have unit norm; $\boldsymbol{x}$ is the discrete token selection indicator vector, K is token budget, and λ balances the maximization of relevance ($\boldsymbol{r}_V^\top \boldsymbol{x}$) against the minimization of redundancy ($\boldsymbol{x}^\top \mathbf{S}\boldsymbol{x}$). To solve the QP, we use the greedy marginal-gain strategy in Algorithm 1. The selected visual tokens are kept in their original patch order and, together with the complete text sequence, fed into the language-model backbone for inference.

Algorithm 1: Greedy Token Selection

Input: relevance $\boldsymbol{r}_V$, redundancy $\mathbf{S}$, token budget K, weight λ
Initialize $\mathcal{I} \leftarrow \emptyset$, $\boldsymbol{p} \leftarrow 0$
for $t = 1$ **to** K **do**
  $i^\star \leftarrow \arg\max_{i \notin \mathcal{I}} (\boldsymbol{r}_{V,i} - 2\lambda \boldsymbol{p}_i)$
  $\mathcal{I} \leftarrow \mathcal{I} \cup \{i^\star\}$
  $\boldsymbol{p} \leftarrow \boldsymbol{p} + \mathbf{S}_{:,i^\star}$
end for
return $\text{sort}(\mathcal{I})$

Computational Complexity. Let m/n be the numbers of textual/visual tokens, with $m \ll n$, and let $\ell = K \ll n$. The adjacency matrix $\mathbf{P}$ requires $\mathcal{O}(m^2 + n^2 + mn)$ time. The diffusion in Eq. (7) costs $\mathcal{O}((m+n)\ell^2 + \ell^3)$, while that in Eq. (10) costs $\mathcal{O}(n^2\ell + n\ell^2)$. The QP-solver in Eq. (13) takes $O(nK)$ time, linear to the number of iterations (Nguyen et al., 2023). Overall, the cost is quadratic with the number of visual tokens, but this is achieved through matrix multiplication highly parallelizable on GPUs, making the procedures very efficient in practice. On average, PCDM introduces **only 7.0 ms additional latency per image** when selecting 160 visual tokens. Detailed runtime breakdown in Appendix E.2. Future work will further investigate the trade-off between efficiency and the quality of the pruning solutions.

4 Experiments

Models and benchmarks. We evaluate on five VLMs from LLaVA and Qwen2.5-VL families: LLaVA-1.5-7B/13B (Liu et al., 2024a), LLaVA-NeXT-7B/13B (Liu et al., 2024b), and Qwen2.5-VL-7B (Bai et al., 2025). We used ten multimodal benchmarks: GQA (Hudson & Manning, 2019), MMBench (MMB) (Liu et al., 2024c), MME (Fu et al., 2026), POPE (Li et al., 2023), ScienceQA-IMG (Lu et al., 2022), VQA$^{\text{v2}}$ (Goyal et al., 2017), TextVQA (VQA$^{\text{Text}}$) (Singh et al., 2019), SEED-Bench (SEED$^{\text{I}}$) (Li et al., 2024), MMVet (Yu et al., 2023), and VizWiz (Gurari et al., 2018).

Experimental settings. In evaluations, **PCDM uses a fixed hyperparameter setting in all cases**: $\alpha = 0.2$, $a = b = 0.2$, $\lambda = 0.05$, and $\gamma = 0.2$. Experiments were run on NVIDIA RTX A6000 GPUs. For competing methods, we used their official implementations and recommended parameter settings if available (marked with * in the results); otherwise, we report results directly from original papers when no code was released, following common practices in the literature (Zou et al., 2026; Fang et al., 2026b; Ma et al., 2026).

4.1 Main Results

Results on LLaVA-1.5-7B. As shown in Table 1, PCDM preserves 100.0%, 99.0%, and 97.5% of the full model's average performance when retaining 192, 128, and 64 visual tokens per image,

Table 1: Performance comparison on 10 benchmarks under different pruning ratios using LLaVA-1.5-7B. The 1st/2nd highest score marked by **bold** / underlines.

Method	MME	MMB	GQA	VQA^{v2}	SQA	VizWiz	VQA^{Text}	POPE	SEED^{I}	MMVet	Avg.
Upper Bound, 576 Tokens **(100%)**											
LLaVA-1.5-7B*	1862	64.7	61.9	78.5	69.5	50.1	58.2	85.9	66.2	31.1	100.0%
Retain 192 Tokens (↓66.7%)											
FastV* (ECCV'24)	1815	63.8	57.8	75.94	69.3	51.22	**57.9**	77.8	63.6	28.8	96.7%
SparseVLM (ICML'25)	1721	62.5	57.6	75.6	69.1	50.0	56.1	83.6	55.8	30.2	95.3%
PDrop (CVPR'25)	1797	63.3	57.3	76.4	69.2	50.0	56.5	84.8	57.2	30.8	96.5%
HiRED (AAAI'25)	1737	62.8	58.7	74.9	68.4	50.1	47.4	82.8	–	–	95.8%
DivPrune (CVPR'25)	1752	62.3	60.0	75.5	68.7	**55.6**	56.4	87.0	58.6	–	98.3%
VisionZip (CVPR'25)	1783	63.0	59.3	76.8	68.9	50.5	57.3	85.3	56.4	31.7	97.2%
PRUNESID (ICLR'26)	1791	63.7	60.1	76.8	68.5	55.4	56.7	86.9	59.0	–	98.9%
CDPruner* (NeurIPS'25)	1786	62.7	60.3	**77.2**	69.0	52.2	57.2	**87.2**	64.7	31.0	98.9%
MMTok* (ICLR'26)	1784	63.6	60.2	77.1	68.8	52.4	57.3	86.8	64.2	30.6	98.8%
SpecFlow (ICML'26)	**1827**	**65.8**	58.3	76.4	69.7	50.5	**57.9**	85.8	–	–	99.3%
HoloV (NeurIPS'25)	1820	65.4	59.0	76.7	69.8	50.9	57.4	85.6	–	–	99.4%
PCDM (Ours)	1792	63.3	**61.4**	**77.2**	**71.0**	52.3	57.5	87.0	**65.5**	**31.9**	**100.0%**
Retain 128 Tokens (↓77.8%)											
FastV* (ECCV'24)	1774	63.4	55.8	73.2	68.1	51.4	56.9	72.1	61.4	28.9	94.5%
SparseVLM (ICML'25)	1696	60.0	56.0	73.8	67.1	51.0	54.9	80.5	53.4	30.0	93.2%
PDrop (CVPR'25)	1761	61.6	57.1	76.0	68.4	49.6	56.6	82.6	56.2	31.0	95.4%
HiRED (AAAI'25)	1710	61.5	57.2	73.4	68.1	51.3	46.1	79.8	–	–	94.3%
DivPrune (CVPR'25)	1752	62.3	59.2	74.7	69.0	55.6	56.0	86.9	57.1	–	97.5%
VisionZip (CVPR'25)	1762	62.0	57.6	75.6	68.9	50.0	56.8	83.2	54.9	**32.6**	96.1%
PRUNESID (ICLR'26)	1749	62.1	58.8	75.3	68.3	**55.8**	54.7	86.5	57.8	–	97.2%
CDPruner* (NeurIPS'25)	1775	62.5	59.8	**76.6**	68.9	52.7	56.3	**87.5**	64.1	29.3	98.0%
MMTok* (ICLR'26)	1747	62.4	59.5	76.4	69.3	53.0	**57.0**	86.3	63.1	31.0	98.3%
SpecFlow (ICML'26)	1794	**64.3**	57.6	75.3	69.9	51.1	56.8	84.9	–	–	98.1%
HoloV (NeurIPS'25)	**1802**	63.9	57.7	75.5	69.8	51.5	56.8	84.0	–	–	98.1%
PCDM (Ours)	1785	61.6	**60.5**	76.5	**70.9**	53.0	56.7	86.8	**64.9**	31.0	**99.0%**
Retain 64 Tokens (↓88.9%)											
FastV* (ECCV'24)	1565	60.1	51.6	66.4	67.9	51.4	55.1	59.6	55.4	26.0	87.7%
SparseVLM (ICML'25)	1505	56.2	52.7	68.2	62.2	49.6	51.8	75.1	51.1	23.3	85.7%
PDrop (CVPR'25)	1561	58.8	47.5	73.3	69.0	50.2	50.6	76.2	53.0	30.5	89.8%
HiRED (AAAI'25)	1599	60.2	54.6	68.7	68.2	50.2	44.2	73.6	–	–	90.8%
DivPrune (CVPR'25)	1638	59.3	57.6	72.9	68.3	**57.5**	55.5	85.6	55.4	–	95.6%
VisionZip (CVPR'25)	1690	60.1	55.1	72.4	69.0	51.9	55.5	77.0	52.2	**31.7**	93.4%
PRUNESID (ICLR'26)	**1733**	58.8	57.1	73.7	67.8	56.9	54.2	83.8	56.1	–	95.5%
CDPruner* (NeurIPS'25)	1709	60.4	58.6	**75.3**	68.4	53.1	55.6	**87.5**	62.4	28.5	96.3%
MMTok* (ICLR'26)	1677	61.6	58.2	75.2	69.3	53.8	55.9	85.8	61.5	26.2	95.6%
SpecFlow (ICML'26)	1713	**63.7**	55.3	73.7	**69.7**	52.4	54.9	80.5	–	–	96.0%
HoloV (NeurIPS'25)	1715	63.3	55.3	72.8	69.5	52.8	55.4	80.3	–	–	96.0%
PCDM (Ours)	1715	60.9	**59.0**	**75.3**	69.2	53.2	**56.2**	87.0	**62.9**	30.7	**97.5%**

Table 2: Performance comparison of LLaVA-NeXT-7B with 160 retained visual tokens, corresponding to a 94.4% pruning ratio. The 1st/2nd highest scores marked by **bold**/underlines.

Method	MME	MMB	GQA	VQA^{v2}	SQA	VizWiz	VQA^{Text}	POPE	SEED^{I}	MMVet	Avg.
Upper Bound, 2880 Tokens **(100%)**											
LLaVA-NeXT-7B*	1856	67.9	64.2	81.9	70.3	57.1	61.2	86.4	70.2	40.5	100.0%
Retain 160 Tokens (↓94.4%)											
FastV* (ECCV'24)	1494	59.9	53.6	67.8	67.8	50.7	54.5	62.4	54.9	28.9	83.1%
SparseVLM (ICML'25)	1542	63.1	51.2	66.3	67.5	–	46.4	–	–	–	88.5%
VisionZip (CVPR'25)	1630	60.1	55.5	71.4	68.3	–	56.2	74.8	58.3	–	88.9%
DivPrune (CVPR'25)	1658	62.2	57.7	73.9	68.0	–	52.4	79.3	62.5	–	90.5%
PRUNESID (ICLR'26)	1704	60.8	58.9	73.8	67.1	–	–	76.9	62.5	–	91.0%
MMTok* (ICLR'26)	1710	**65.8**	**62.1**	74.6	68.4	51.6	**59.4**	83.7	64.1	29.9	92.3%
CDPruner* (NeurIPS'25)	1724	63.4	60.6	**76.5**	67.5	54.9	55.5	87.0	65.7	32.7	93.2%
PCDM (Ours)	**1783**	63.6	61.3	76.2	**69.4**	**55.9**	56.4	**87.1**	**66.3**	**33.6**	**94.5%**

respectively. With a 192-token budget, PCDM nearly matches the full model performance. Even with only 64 visual tokens (88.9% token reduction), PCDM achieves the highest average relative performance of 97.5% among all competitors.

Table 3: Relative average accuracy (Avg.%, ↑) over 10 benchmarks at three pruning ratios for LLaVA-1.5 and LLaVA-NeXT series models.

Method	LLaVA-1.5-7B			LLaVA-1.5-13B			LLaVA-NeXT-7B			LLaVA-NeXT-13B		
	576 tokens			576 tokens			Upper (Up.) 2880 tokens			Upper (Up.) 2880 tokens		
Compress Ratio	↓67%	↓78%	↓89%	↓67%	↓78%	↓89%	↓78%	↓89%	↓94%	↓78%	↓89%	↓94%
Remain Token	192	128	64	192	128	64	Up., 640	Up., 320	Up., 160	Up., 640	Up., 320	Up., 160
FastV*	96.7%	94.5%	87.7%	97.3%	95.3%	90.9%	95.1%	90.0%	83.1%	96.3%	92.7%	84.1%
CDPruner*	98.9%	98.0%	96.3%	97.5%	97.5%	96.3%	97.6%	95.7%	93.2%	96.6%	94.8%	92.9%
MMTok*	98.8%	98.3%	95.6%	98.1%	96.9%	95.9%	97.1%	94.4%	92.3%	96.2%	94.6%	92.9%
PCDM (Ours)	**100.0%**	**99.0%**	**97.5%**	**98.4%**	**97.9%**	**97.2%**	**98.4%**	**96.5%**	**94.5%**	**97.6%**	**95.4%**	**93.7%**

Results on LLaVA-NeXT-7B. On LLaVA-NeXT-7B with high-resolution inputs, PCDM achieves the best average relative performance at all three token budgets, retaining 98.4%, 96.5%, and 94.5% of the full model's average performance with 640, 320, and 160 tokens, respectively (Table 3). Under the most aggressive 160-token setting, corresponding to a 94.4% pruning ratio, it outperforms the strongest baseline, CDPruner, by 1.3% (Table 2). Notably, reducing the token budget from 640 to 160 lowers its average relative performance by only 3.9%, indicating graceful degradation and a superior performance–efficiency trade-off under high-resolution inputs and aggressive pruning.

Results on Qwen2.5-VL-7B. To examine the generalizability of PCDM beyond LLaVA, we further evaluate it on Qwen2.5-VL-7B. As shown in Table 4, PCDM preserves 98.5%, 96.3%, and 91.0% of the full model's average performance at pruning ratios of 66.7%, 77.8%, and 88.9%, respectively, consistently outperforming all competitors. Notably, its improvement over the strongest baseline increases from 1.2% to 4.4% as the pruning ratio rises from 66.7% to 88.9%, a widening margin that is more pronounced than that observed on LLaVA (0.3% to 1.2%). The consistent gains under increasingly aggressive pruning across both model families demonstrate that PCDM generalizes across MLLM architectures and remains robust under highly constrained token budgets.

Table 4: Qwen2.5-VL-7B – different pruning ratios.

Method	MMB	POPE	SQA	VQA^{Text}	MMVet	Avg.
Upper Bound **(100%)**						
Qwen2.5-VL-7B	82.74	86.40	88.54	82.87	62.20	100.0%
Token Pruning Rate (↓66.7%)						
FastV* (ECCV'24)	79.76	83.38	86.06	77.90	58.70	95.7%
CDPruner* (NeurIPS'25)	80.53	83.69	88.07	80.29	59.70	97.3%
MMTok* (ICLR'26)	80.87	84.55	88.16	78.05	59.00	96.8%
PCDM (Ours)	**82.31**	**84.61**	**88.28**	**81.29**	**60.40**	**98.5%**
Token Pruning Rate (↓77.8%)						
FastV* (ECCV'24)	76.02	80.40	84.97	69.00	47.50	88.1%
CDPruner* (NeurIPS'25)	78.15	82.40	87.22	77.15	54.50	93.8%
MMTok* (ICLR'26)	78.74	82.40	87.69	75.00	53.50	93.2%
PCDM (Ours)	**79.08**	**83.33**	**88.07**	**80.58**	**57.70**	**96.3%**
Token Pruning Rate (↓88.9%)						
FastV* (ECCV'24)	60.12	71.52	82.80	60.30	40.70	77.4%
CDPruner* (NeurIPS'25)	72.28	78.80	86.21	67.79	46.80	86.6%
MMTok* (ICLR'26)	72.57	79.68	86.44	64.87	42.00	84.7%
PCDM (Ours)	**72.96**	**82.70**	**87.01**	**71.02**	**54.20**	**91.0%**

Results Across Model Scales. Table 3 compares PCDM on LLaVA-1.5-7B/13B and LLaVA-NeXT-7B/13B. PCDM achieves the highest average relative performance in all 12 model–budget settings, showing that its advantages extend consistently across model scales and input resolutions.

Overall, the gains of PCDM over competing methods are consistent across base LLMs, benchmark datasets, token pruning ratios, and model scales, indicating robust performance in diverse settings.

4.2 Ablation Studies

Table 5 examines three key design choices of PCDM on LLaVA-1.5-7B with 64 visual tokens. *Why preserve position and content as distinct relational structures but not directly mix them?* We replace PCDM with a variant that directly sums the position and content relation matrices (*Shared-node fusion*). Performance drops by 2.1%, i.e., mixing where-what relations–even at the relational level–loses useful structure. This supports preserving position/content separately before they should interact. (2) *Is higher-order token relation beneficial?* Removing the globally diffused relational matrix $\widetilde{\boldsymbol{S}}^{\text{global}}$ and retaining only local version $\mathbf{S}^{\text{ori}}$ causes a 0.9% drop, indicating that multi-hop affinities provide useful cues beyond local relations. (3) *Is redundancy-aware selection necessary?* Removing the redundancy penalty and selecting tokens solely by relevance scores reduces performance by 3.4% on average (5.9% on TextVQA). Appendix C further shows that traditional position or content redundancy measure is less effective than joint, high-order *where-what* modeling in PCDM.

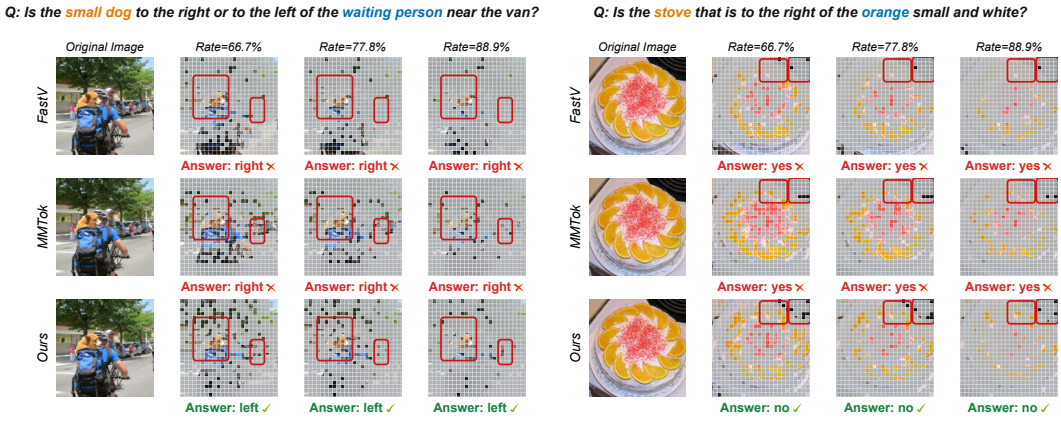


Figure 3: Qualitative examples of visual tokens retained by FastV, MMTok, and PCDM under different pruning ratios (correct answers in green and wrong answers in red).

Figure 4 examines effect of modeling progressively higher-order token relations as redundancy measure. As the number of random-walk hops increases, performance improves consistently, showing that global, higher-order relational reasoning is useful for token pruning besides local interactions.

Table 5: Ablation study of PCDM.

Variants	SQA	MMB	POPE	VQA^{Text}	Avg.
LLaVA-1.5-7B	69.5	64.7	85.9	58.2	100.0%
Shared-node fusion	68.3	59.7	85.7	54.1	95.8%
Local redundancy only	69.0	60.5	86.2	55.2	97.0%
Relevance only	68.1	60.3	86.4	50.3	94.5%
PCDM (Ours)	**69.2**	**60.9**	**87.0**	**56.2**	**97.9%**

Table 6: Efficiency analysis of PCDM.

Method	Token	Infer. (ms)↓	Prefill (ms)↓	Mem. (GB)↓	Acc.↑
LLaVA-NeXT	2880	243.4	211.5	17.4	70.3
FastV* (ECCV'24)	160	180.4	91.3	16.4	67.8
CDPruner* (NeurIPS'25)	160	98.2	70.1	14.8	67.5
MMTok* (ICLR'26)	160	102.2	71.3	**14.1**	68.3
PCDM (Ours)	160	**98.1**	**69.4**	14.6	**69.4**

Figure 3 visualizes the selected visual tokens for a few methods. As the pruning ratio increases from 66.7% to 88.9%, PCDM consistently preserves key visual details and yields correct answers even at high pruning ratio. This demonstrates that under aggressive pruning, PCDM effectively filters out redundant information while better preserving complementary and query-relevant visual evidence. **Note that token importance from VLM need not necessarily align with human visual judgement**: a model may still answer correctly from a sparse and seemingly incomplete subset of image patches. Therefore the selected regions should not be interpreted solely through human-perceived saliency. More examples are in Appendix H.

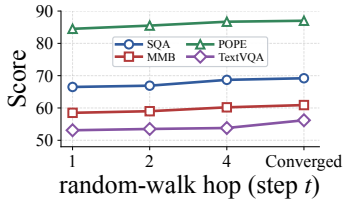


Figure 4: Higher-order token affinities lead to better results.

Table 6 evaluates PCDM efficiency on ScienceQA using LLaVA-NeXT-7B. Retaining only 160 visual tokens, PCDM achieves a 2.48× end-to-end speedup and 3.05× prefill speedup, reduces peak GPU memory by 16.1%, and preserves 98.7% of the original performance. It has higher accuracy and lower latency than competitors; although MMTok uses slightly less memory, PCDM offers a better overall efficiency–performance trade-off. Importantly, these measurements include PCDM's own overhead, reflecting actual end-to-end gains rather than theoretical FLOP reductions alone.

5 Conclusion

In this work, we proposed PCDM, a training-free visual token pruning method that models heterogeneous attribute relations across modalities for efficient VLM inference. PCDM preserves distinct structural properties while enabling flexible interaction, providing valuable high-order relational signals for selecting task-relevant yet non-redundant visual tokens. Experiments demonstrate promising accuracy–efficiency trade-offs. In the future we shall further improve the sampling procedure to recover the inner structures of the global relational system, and derive thorough analysis.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9392–9401. IEEE, 2025a.

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models, 2025b. URL https://arxiv.org/abs/2503.02175.

Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S. Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models, 2024. URL https://arxiv.org/abs/2408.10945.

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.

Jingyang Chen, Ping Li, Jiancheng Lv, Hongyuan Zha, Kai Zhang, and Jie Zhang. Learning temporal features with alternated similarity and proximity attention for time-series prediction. *IEEE Transactions on Neural Networks and Learning Systems*, 36(9):16339–16350, 2025. doi: 10.1109/TNNLS.2025.3559222.

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, 2024.

Tzu-Chun Chien, Chieh-Kai Lin, Shiang-Feng Tsai, Ruei-Chi Lai, Hung-Jen Chen, and Min Sun. Grounding-aware token pruning: Recovering from drastic performance drops in visual grounding caused by pruning. *arXiv preprint arXiv:2506.21873*, 2025.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

M Donello, G Palkar, MH Naderi, DC Del Rey Fernández, and H Babaee. Oblique projection for scalable rank-adaptive reduced-order modelling of nonlinear stochastic partial differential equations with time-dependent bases. *Proceedings of the Royal Society A*, 479(2278):20230320, 2023.

Sixun Dong, Juhua Hu, Mian Zhang, Ming Yin, Yanjie Fu, and Qi Qian. Mmtok: Multimodal coverage maximization for efficient inference of vlms. *arXiv preprint arXiv:2508.18264*, 2025.

Yuxiang Duan, Ao Li, Yingqin Li, Luyu Li, and Pengwei Wang. Gridprune: From "where to look" to "what to select" in visual token pruning for mllms. *Neurocomputing*, pp. 133887, 2026.

Zhengyao Fang, Pengyuan Lyu, Chengquan Zhang, Guangming Lu, Jun Yu, and Wenjie Pei. Prune redundancy, preserve essence: Vision token compression in VLMs via synergistic importance-diversity. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=i36E5Ezm0H.

Zhengyao Fang, Pengyuan Lyu, Chengquan Zhang, Guangming Lu, Jun Yu, and Wenjie Pei. Prune redundancy, preserve essence: Vision token compression in vlms via synergistic importance-diversity, 2026b. URL https://arxiv.org/abs/2603.09480.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

Mouxiao Huang, Borui Jiang, Dehua Zheng, Hailin Hu, Kai Han, and Xinghao Chen. Ppe: Positional preservation embedding for token compression in multimodal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 67656–67676, 2026a.

Yihong Huang, Fei Ma, Yihua Shao, Jingcai Guo, Zitong YU, Laizhong Cui, and Qi Tian. Nüwa: Mending the spatial integrity torn by VLM token pruning. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=C9yclwdquU.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Ahmadreza Jeddi, Negin Baghbanzadeh, Elham Dolatabadi, and Babak Taati. Similarity-aware token pruning: Your vlm but faster. *arXiv preprint arXiv:2503.11549*, 2025.

Bin Kang, Bin Chen, Junjie Wang, Yulin Li, Junzhi Zhao, Junle Wang, and Zhuotao Tian. Calibclip: Contextual calibration of dominant semantics for text-driven image retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5140–5149, 2025.

Guolin Ke, Di He, and Tie-Yan Liu. Rethinking positional encoding in language pre-training. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=09-528y2Fgf.

Gary Kochenberger, Jin-Kao Hao, Fred Glover, Mark Lewis, Zhipeng Lü, Haibo Wang, and Yang Wang. The unconstrained binary quadratic programming problem: a survey. *Journal of combinatorial optimization*, 28(1):58–81, 2014.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Zhaoyang Li, Yanjun Li, Wangkai Li, Yujia Chen, and Tianzhu Zhang. Spectral heat flow for conservative token condensation in vision-language models. *arXiv preprint arXiv:2607.10640*, 2026.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, january 2024. *URL https://llava-vl.github. io/blog/2024-01-30-llava-next*, 1(8), 2024b.

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

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Qiyanhui Lu, Han Wu, Rongjian Xu, Tingzhang Luo, Cheng Fan, Xinghao Chen, Minjing Dong, Jufeng Yang, and Jianyuan Guo. Rora: Role-oriented regional allocation for visual token pruning in mllms. *arXiv preprint arXiv:2608.07088*, 2026.

Kexin Ma, Jing Xiao, Chaofeng Chen, Geyong Min, Guibo Zhu, Jinqiao Wang, and Liang Liao. Decoupled similarity for task-aware token pruning in large vision-language models. *arXiv preprint arXiv:2604.11240*, 2026.

Duy Nguyen, Ngoc Bui, and Viet Anh Nguyen. Feasible recourse plan via diverse interpolation. In *International Conference on Artificial Intelligence and Statistics*, pp. 4679–4698. PMLR, 2023.

Yixian Shen, Zhiheng Yang, Qi Bi, Changshuo Wang, Shuai Wang, Jia-Hong Huang, George Floros, Prayag Tiwari, and Anuj Pathania. Spectral-progressive thought flow for lightweight multimodal reasoning, 2026. URL https://arxiv.org/abs/2606.02842.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Zhichao Sun, Yidong Ma, Gang Liu, Yibo Chen, Xu Tang, Yao Hu, and Yongchao Xu. Ivc-prune: Revealing the implicit visual coordinates in lvlms for vision token pruning. *arXiv preprint arXiv:2602.03060*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction, 2025. URL https://arxiv.org/abs/2410.17247.

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.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, Minqi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. [cls] attention is all you need for training-free visual token pruning: Make vlm inference faster. *arXiv e-prints*, pp. arXiv–2412, 2024a.

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *Advances in Neural Information Processing Systems*, 38:25438–25468, 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*, 2024b.

Xin Zou, Di Lu, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Xu Zheng, Linfeng Zhang, and Xuming Hu. Don't just chase "highlighted tokens" in mllms: Revisiting visual holistic context retention. *Advances in Neural Information Processing Systems*, 38:39800–39832, 2026.

A AI USE STATEMENT

Large language models were used to assist with English-language editing of the manuscript, including grammatical correction, wording refinement, and sentence restructuring to improve overall readability. All research ideas, methodological contributions, experimental evaluations, analyses, and conclusions presented in this work were independently developed and verified by the authors.

B ADDITIONAL METHODOLOGICAL DETAILS

This appendix derives the low-rank inverse used in the text-guided relevance diffusion of Sec. 3.2 and describes the type-aware landmark selection strategy used to construct the CUR approximation. All notation follows the definitions in the main text.

B.1 CUR–WOODBURY DERIVATION

Let $d = 2m + 2n$ denote the total number of nodes in the global PCDM system, ordered as $[U_c, U_p, V_c, V_p]$. We denote the corresponding column-normalized transition matrix by $\widetilde{\mathbf{P}} \in \mathbb{R}^{d\times d}$. Given the initial distribution $\boldsymbol{r}^{(0)}$, the random walk with restart is iteratively defined as

$$\boldsymbol{r}^{(t)} = (1-\alpha)\widetilde{\mathbf{P}}\boldsymbol{r}^{(t-1)} + \alpha\boldsymbol{r}^{(0)}. \tag{14}$$

At convergence, $\boldsymbol{r}^{(t)} = \boldsymbol{r}^{(t-1)} = \boldsymbol{r}^{\star}$, yielding

$$\boldsymbol{r}^{\star} = (1-\alpha)\widetilde{\mathbf{P}}\boldsymbol{r}^{\star} + \alpha\boldsymbol{r}^{(0)}. \tag{15}$$

Rearranging the terms gives

$$\left[\mathbf{I}_{2m+2n} - (1-\alpha)\widetilde{\mathbf{P}}\right]\boldsymbol{r}^{\star} = \alpha\boldsymbol{r}^{(0)}, \tag{16}$$

and hence the steady-state distribution is

$$\boldsymbol{r}^{\star} = \alpha\left[\mathbf{I}_{2m+2n} - (1-\alpha)\widetilde{\mathbf{P}}\right]^{-1}\boldsymbol{r}^{(0)}. \tag{17}$$

This recovers the closed-form solution in Eq. (4). Replacing $\widetilde{\mathbf{P}}$ with its CUR approximation from Eq. (6) gives the following low-rank form:

$$\boldsymbol{r}^{\star} \approx \alpha\left[\mathbf{I}_{2m+2n} - (1-\alpha)\mathbf{C}\mathbf{U}^{\dagger}\mathbf{R}\right]^{-1}\boldsymbol{r}^{(0)}. \tag{18}$$

Although the CUR approximation expresses the transition matrix in a low-rank form, directly evaluating Eq. (18) still requires inverting a $(2m + 2n) \times (2m + 2n)$ matrix. We therefore exploit the low-rank structure induced by CUR and apply the Woodbury identity:

$$(\mathbf{I} - \mathbf{A}\mathbf{B})^{-1} = \mathbf{I} + \mathbf{A}(\mathbf{I} - \mathbf{B}\mathbf{A})^{-1}\mathbf{B}, \quad \mathbf{A} = (1-\alpha)\mathbf{C}, \quad \mathbf{B} = \mathbf{U}^{\dagger}\mathbf{R}, \tag{19}$$

where $\mathbf{A} \in \mathbb{R}^{(2m+2n)\times\ell}$ and $\mathbf{B} \in \mathbb{R}^{\ell\times(2m+2n)}$. Under this substitution, $\mathbf{A}\mathbf{B} = (1-\alpha)\mathbf{C}\mathbf{U}^{\dagger}\mathbf{R}$ corresponds exactly to the low-rank correction of $(1-\alpha)\mathbf{C}\mathbf{U}^{\dagger}\mathbf{R}$ in Eq. (18), whereas the reverse product $\mathbf{B}\mathbf{A} = (1-\alpha)\mathbf{U}^{\dagger}\mathbf{R}\mathbf{C}$ lies in the much smaller ℓ-dimensional space. We therefore define

$$\mathbf{H} = \mathbf{I}_{\ell} - \mathbf{B}\mathbf{A} = \mathbf{I}_{\ell} - (1-\alpha)\mathbf{U}^{\dagger}\mathbf{R}\mathbf{C}, \tag{20}$$

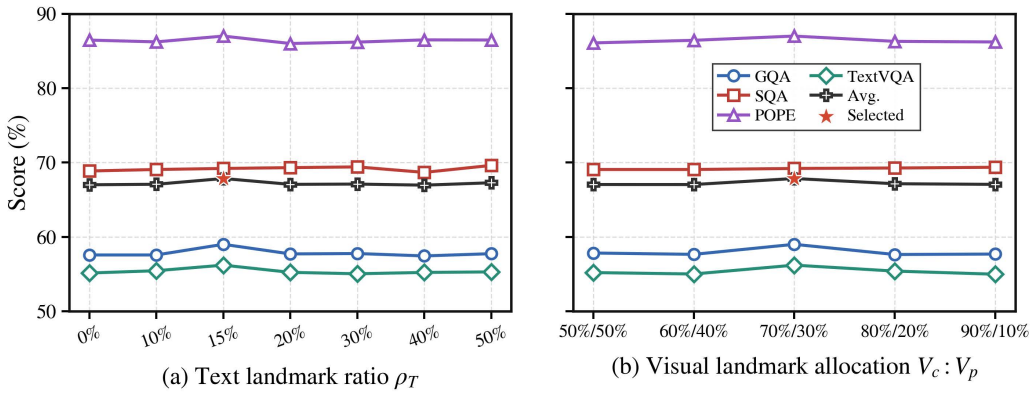
where $\mathbf{H} \in \mathbb{R}^{\ell\times\ell}$. Applying Eq. (19) and Eq. (20) to the original inverse term in Eq. (18) gives

$$\begin{aligned}\left[\mathbf{I}_{2m+2n} - (1-\alpha)\mathbf{C}\mathbf{U}^{\dagger}\mathbf{R}\right]^{-1} &= \mathbf{I}_{2m+2n} + (1-\alpha)\mathbf{C}\left[\mathbf{I}_{\ell} - (1-\alpha)\mathbf{U}^{\dagger}\mathbf{R}\mathbf{C}\right]^{-1}\mathbf{U}^{\dagger}\mathbf{R} \\ &= \mathbf{I}_{2m+2n} + (1-\alpha)\mathbf{C}\mathbf{H}^{-1}\mathbf{U}^{\dagger}\mathbf{R}.\end{aligned} \tag{21}$$

Substituting Eq. (21) into Eq. (18) yields the efficient low-rank form

$$\boldsymbol{r}^{\star} \approx \alpha\left[\mathbf{I}_{2m+2n} + (1-\alpha)\mathbf{C}\mathbf{H}^{-1}\mathbf{U}^{\dagger}\mathbf{R}\right]\boldsymbol{r}^{(0)}. \tag{22}$$

Thus, the CUR decomposition and the Woodbury identity jointly reduce the matrix inversion from size $(2m+2n)\times(2m+2n)$ to $\ell\times\ell$. Since $\ell \ll 2m+2n$, this substantially reduces the computational cost of evaluating the steady-state solution.



(a) Text landmark ratio ρ_T

(b) Visual landmark allocation $V_c : V_p$

Figure 5: PCDM landmark budget sensitivity analysis.

B.2 Landmark Selection

CUR uses the rows and columns indexed by a small set of landmark nodes to approximate the full transition matrix. The PCDM graph contains four node types: text-content nodes U_c, text-position nodes U_p, visual-content nodes V_c, and visual-position nodes V_p. Landmarks should cover these node types to preserve query semantics, cross-modal relations, and spatial structure. Selecting visual-content landmarks solely based on query relevance may cause them clustered in a few regions, while uniform sampling may overlook query-relevant cross-modal connections. We therefore first allocate the landmark budget and then select representative nodes within each type.

Landmark budget allocation. Let l denote the total landmark budget, and let l_T and l_V denote the text-side and visual-side budgets, respectively, with $l = l_T + l_V$. We set $l = K$, where K is the number of retained visual tokens. A fraction ρ_T of the budget is assigned to text nodes and is approximately divided equally between U_c and U_p. The remaining budget is assigned to visual nodes, with 70% allocated to V_c and the remaining 30% allocated to V_p. The numbers are rounded to integers while preserving the total budget.

We evaluate different budget allocations on LLaVA-1.5-7B with $l = K = 64$ on GQA, ScienceQA, POPE, and TextVQA. As shown in Fig. 5, too few text landmarks weaken the preservation of query conditions, whereas too many reduce the capacity available for visual content and spatial structure. A text-landmark ratio of $\rho_T = 15\%$ achieves the best average performance. On the visual side, allocating too few landmarks to V_c weakens the representation of query-relevant semantics, while allocating too few to V_p reduces spatial coverage. Allocating 70% of the visual budget to V_c and 30% to V_p provides a favorable balance.

We therefore use $\rho_T = 15\%$ in the main experiments, with 70% and 30% of the visual budget assigned to V_c and V_p, respectively. When $l = K = 64$, this gives approximately 10 text landmarks, including about 5 from each of U_c and U_p, and 54 visual landmarks, including about 38 from V_c and 16 from V_p.

Landmark node selection. The budget allocation determines how many landmarks are selected from each node type. We then select specific nodes within their respective budgets.

For text nodes, we compute the inner product between each text-token representation and the mean-pooled query representation. The content nodes corresponding to the highest-scoring tokens are selected as U_c landmarks, together with their corresponding position nodes as U_p landmarks. This paired selection preserves both query semantics and text-position information.

For visual-content nodes, we combine direct query relevance with one-step cross-modal support. Following the main text, text and visual token representations are ℓ_2-normalized, and the pooled query representation is normalized in the same way. The inner-product score for the i-th visual token is $a_i = \langle \bar{\boldsymbol{v}}_i, \bar{\boldsymbol{u}}_q \rangle$. We apply softmax to all direct relevance scores to obtain the normalized relevance scores d_i.

Table 7: Ablation study of landmark selection strategies on LLaVA-1.5-7B with $l = K = 64$.

Selection strategy	GQA	SQA	POPE	TextVQA	Avg.
Visual-content landmark scoring					
Only direct relevance	57.88	68.91	86.60	55.25	97.10%
Only one-step propagation response	57.80	68.72	86.13	55.19	96.83%
Direct relevance + one-step response	**59.00**	**69.20**	**87.00**	**56.20**	**98.18%**
Visual-position landmark selection					
Paired selection only	57.54	68.96	86.27	55.14	96.84%
Uniform sampling only	57.64	69.01	86.39	55.57	97.12%
50% paired + 50% uniform	**59.00**	**69.20**	**87.00**	**56.20**	**98.18%**

We then propagate the initial text-query distribution for one step on the full PCDM graph and retain the responses corresponding to V_c. These responses are sum-normalized over the visual-content nodes to obtain the one-step propagation scores $\widehat{h}_i^{(1)}$. We average d_i and $\widehat{h}_i^{(1)}$ to obtain the final visual-content landmark score s_i^{lm}. The visual-content nodes with the highest scores are selected within the allocated V_c budget.

For visual-position nodes, 50% of the budget is assigned to the position nodes corresponding to the selected visual-content landmarks. The remaining 50% is sampled uniformly over the visual grid. This preserves content–position correspondence while maintaining global spatial coverage.

The final landmark set consists of landmarks from U_c, U_p, V_c, and V_p. Given this set, we extract the corresponding rows, columns, and intersection submatrix from the column-normalized transition matrix $\widetilde{\mathbf{P}}$ to construct the CUR approximation.

Adaptation to Qwen2.5-VL. We apply the same landmark construction and selection rules to Qwen2.5-VL, using its visual token sequence and corresponding two-dimensional grid to instantiate V_c and V_p. No model-specific ratios or selection rules are introduced, and the selected visual tokens are restored to their original sequence order before inference.

Landmark selection ablation. After determining the budget allocation, we compare different selection strategies for visual-content and visual-position landmarks.

As shown in Table 7, direct relevance and one-step propagation provide complementary information. Using only direct relevance does not exploit graph-based cross-modal support, whereas using only the propagation response lacks the direct match between visual tokens and the overall query. Combining both scores achieves the highest average retention.

For visual-position nodes, uniform sampling alone weakens content–position correspondence, while paired selection alone concentrates landmarks in a few query-related regions. The 50% paired and 50% uniform strategy provides a better balance between correspondence and global spatial coverage.

C Analysis of Redundancy Modeling

To examine the role of higher-order relations in redundancy-aware token selection in more detail, we replace PCDM's redundancy modeling, which combines first-order and higher-order token relations, with variants using only first-order relations. Specifically, we remove the $\widetilde{\mathbf{S}}^{\mathrm{global}}$ term and replace $\mathbf{S}^{\mathrm{ori}}$ in Eq. (11) with three alternatives: position-only redundancy ($\mathbf{K}^{\text{p-p}}_{n\times n}$), defined by spatial proximity on the native 2D visual grid; content-only redundancy ($\mathbf{K}^{\text{c-c}}_{n\times n}$), defined by pairwise inner-product similarity between visual-token representations; and an equal-weight fusion of the normalized position and content matrices ($0.5\mathbf{K}^{\text{p-p}}_{n\times n} + 0.5\mathbf{K}^{\text{c-c}}_{n\times n}$). All other components, including relevance scores, global diffusion modulation, and the QP objective, remain unchanged.

As shown in Table 8, position-only and content-only redundancy modeling based on first-order token relations, as well as their direct combination, all underperform PCDM. Position-only redundancy modeling causes a larger performance drop than content-only modeling, particularly on TextVQA

Table 8: Analysis of redundancy modeling strategies on LLaVA-1.5-7B with $l = K = 64$.

Redundancy modeling	SQA	MMB	POPE	TextVQA	Avg.
Position-only (pairwise spatial proximity)	67.5	59.3	85.7	51.2	94.1%
Content-only (pairwise inner products)	68.2	60.2	86.1	54.4	96.2%
Content + Position (fusion of 1st-order relations)	68.7	60.5	86.4	54.6	96.7%
PCDM (joint, high-order position-content relations)	**69.2**	**60.9**	**87.0**	**56.2**	**97.9%**

(51.2 vs. 54.4), suggesting that content relations provide a stronger cue for redundancy estimation. Nevertheless, both variants, as well as their direct combination, remain inferior to PCDM.

These results indicate that effective redundancy modeling benefits primarily from jointly capturing position–content dependencies and their higher-order structures. While combining content similarity with spatial proximity improves over either cue alone, simple linear fusion still treats them largely as separate first-order relations and does not explicitly model their interactions. PCDM instead propagates relational signals through random walks over the cross-modal position–content manifolds, thereby capturing higher-order dependencies between the two attributes and consistently outperforming all first-order variants. This higher-order joint modeling more reliably distinguishes truly redundant content from spatially complementary evidence.

D Model Architectures and Benchmark Coverage

D.1 Model Architectures

Our evaluation considers five configurations from three representative VLM families, covering fixed-resolution, tiled high-resolution, and native dynamic-resolution visual encoding. This diversity enables a systematic assessment of PCDM across model scales and visual-token layouts.

LLaVA-1.5. LLaVA-1.5 (Liu et al., 2024a) adopts CLIP ViT-L/14-336 as its visual encoder and maps the resulting features into the Vicuna embedding space through a two-layer MLP projector. We use both the 7B and 13B variants. Under its fixed 336×336 input resolution, each image is represented by a 24×24 grid comprising 576 visual tokens.

LLaVA-NeXT. LLaVA-NeXT (LLaVA-1.6) (Liu et al., 2024b) extends LLaVA-1.5 with the AnyRes strategy, which jointly encodes a global view and multiple local image views to preserve fine-grained visual details. Consequently, the number of visual tokens varies dynamically across inputs according to the image aspect ratio and AnyRes grid configuration, reaching up to 2,880 tokens per input.

Qwen2.5-VL. Qwen2.5-VL (Bai et al., 2025) employs a native dynamic-resolution Vision Transformer with window attention, producing visual sequences whose lengths adapt to the input resolution. We use Qwen2.5-VL-7B to evaluate the transferability of PCDM beyond the fixed-grid and tiled encoding schemes of the LLaVA family.

D.2 Benchmark Coverage

The evaluation covers altogether ten benchmarks spanning general visual question answering, compositional reasoning, knowledge- and text-intensive understanding, object hallucination, and integrated multimodal capabilities, which we introduce below.

General visual question answering. VQAv2 (Goyal et al., 2017) assesses open-ended visual question answering using complementary image pairs designed to reduce language priors. GQA (Hudson & Manning, 2019) focuses on compositional reasoning over objects, attributes, and relations through scene-graph-derived questions. VizWiz (Gurari et al., 2018) contains real-world images and questions collected from blind users, emphasizing robustness to degraded visual quality, ambiguity, and unanswerable queries.

Table 9: Per-sample latency breakdown of the PCDM token-selection procedure on ScienceQA examples with 160 retained visual tokens.

Stage	Main operation	Latency (ms) ↓
Graph construction	Heterogeneous relation modeling	3.1
Restart initialization	Query-conditioned restart distribution	0.4
Relation diffusion	CUR-based random walk with restart	2.4
Token selection	QP-based greedy selection	1.1
Total	**PCDM selection overhead**	**7.0**

Knowledge- and text-intensive understanding. ScienceQA (SQA) (Lu et al., 2022) evaluates multiple-choice scientific reasoning that requires the joint use of visual evidence and domain knowledge. TextVQA (VQA^{Text}) (Singh et al., 2019) measures the ability to recognize and reason over text embedded in natural images, thereby requiring the preservation of fine-grained visual evidence.

Object hallucination. POPE (Li et al., 2023) evaluates visual grounding through binary object-presence questions. It measures whether model predictions remain consistent with image content rather than relying on language priors.

Comprehensive multimodal evaluation. MME (Fu et al., 2026) covers perception and cognition through 14 subtasks, including recognition, counting, spatial understanding, OCR, and reasoning. MMBench (Liu et al., 2024c) provides a fine-grained capability taxonomy and systematically constructed multiple-choice questions for multimodal evaluation. SEEDBench-Image (SEED^{I}) (Li et al., 2024) assesses a range of image-understanding capabilities using human-annotated multiple-choice questions. MMVet (Yu et al., 2023) focuses on the integration of recognition, OCR, knowledge, spatial reasoning, mathematics, and language generation in open-ended multimodal tasks.

E Implementation Details

E.1 Hyperparameter Settings

Unless otherwise specified, we use the same hyperparameter setting across all model architectures and benchmarks:

$$\alpha = 0.2, \qquad \ell = K, \qquad a = b = 0.2, \qquad \lambda = 0.05, \qquad \gamma = 0.2. \tag{23}$$

Here, α is the restart probability; ℓ is the number of landmark nodes, set equal to the retained-token budget K; λ controls the trade-off between relevance and redundancy in the QP objective; and γ controls the contribution of global diffusion redundancy.

The coefficients a and b control the content–position balance in relevance estimation and redundancy modeling, respectively. In Eq. (9) , the visual-content and visual-position responses are weighted by a and $1 - a$. In Eq. (10) and Eq. (11) , the content-based and position-based affinities are weighted by $1 - b$ and b. Therefore, setting $a = b = 0.2$ assigns a weight of 0.8 to positional evidence in relevance estimation and a weight of 0.8 to content-based affinity in redundancy estimation. This complementary configuration emphasizes text-conditioned spatial evidence when estimating token relevance, while primarily suppressing semantic overlap during redundancy-aware selection.

E.2 PCDM Cost Breakdown

To further analyze the practical computational overhead of PCDM, we profile the token-selection stage after obtaining visual and textual representations and before VLM prefilling. Experiments are conducted on ScienceQA examples with LLaVA-NeXT-7B, where the retained visual-token budget is fixed at 160. As shown in Table 9, we measure the average per-sample latency of each component in PCDM, including heterogeneous graph construction, restart distribution initialization, relation diffusion, and QP-based greedy token selection.

The results show that the major overhead of PCDM comes from heterogeneous graph construction and relation diffusion, which require 3.1 ms and 2.4 ms, respectively. These costs mainly arise from explicitly modeling multi-attribute relations and capturing global dependencies through high-order

Table 10: Performance comparison on LLaVA-NeXT-7B. The 1st/2nd highest score marked by **bold** / underlines.

Method	MME	MMB	GQA	VQA^{v2}	SQA	VizWiz	VQA^{Text}	POPE	SEED^{I}	MMVet	Avg.
Upper Bound, 2880 Tokens **(100%)**											
LLaVA-NeXT-7B*	1856	67.9	64.2	81.9	70.3	57.1	61.2	86.4	70.2	40.5	100.0%
Retain 640 Tokens (↓77.8%)											
FastV* (ECCV'24)	1686	63.3	60.8	77.1	65.9	55.7	58.1	84.4	67.8	39.5	95.1%
CDPruner* (NeurIPS'25)	1788	65.8	**62.6**	**79.8**	67.9	55.6	58.8	87.3	69.0	39.8	97.6%
MMTok* (ICLR'26)	**1827**	66.2	62.5	76.9	68.0	54.7	**60.4**	86.6	**69.6**	37.8	97.1%
PCDM (Ours)	1816	**66.6**	62.4	**79.8**	**68.5**	**57.2**	60.1	**87.9**	68.1	**40.2**	**98.4%**
Retain 320 Tokens (↓88.9%)											
FastV* (ECCV'24)	1535	61.8	55.1	76.6	67.4	53.3	56.5	78.3	62.8	34.7	90.0%
CDPruner* (NeurIPS'25)	1752	65.1	61.5	**78.4**	67.5	55.2	57.7	87.3	67.4	**37.2**	95.7%
MMTok* (ICLR'26)	**1786**	**66.2**	**62.7**	75.8	67.3	52.5	**59.8**	85.4	65.8	33.1	94.4%
PCDM (Ours)	1766	**66.2**	61.8	78.3	**67.6**	**57.0**	58.0	**87.8**	**68.0**	36.9	**96.5%**
Retain 160 Tokens (↓94.4%)											
FastV* (ECCV'24)	1494	59.9	53.6	67.8	67.8	50.7	54.5	62.4	54.9	28.9	83.1%
CDPruner* (NeurIPS'25)	1724	63.4	60.6	**76.5**	67.5	54.9	55.5	87.0	65.7	32.7	93.2%
MMTok* (ICLR'26)	1710	**65.8**	**62.1**	74.6	68.4	51.6	**59.4**	83.7	64.1	29.9	92.3%
PCDM (Ours)	**1783**	63.6	61.3	76.2	**69.4**	**55.9**	56.4	**87.1**	**66.3**	**33.6**	**94.5%**

propagation. In contrast, restart distribution initialization and QP-based greedy token selection introduce only minor overhead. Overall, PCDM introduces only 7.0 ms additional latency for token selection, while exploiting content-position dual-manifold interactions for token pruning with limited practical overhead.

F Complete Empirical Results

F.1 LLaVA-NeXT-7B

Table 10 shows that on the LLaVA-NeXT-7B model, PCDM achieves the highest average performance retention across all three visual token budgets. Specifically, with 640, 320, and 160 visual tokens, PCDM retains 98.4%, 96.5%, and 94.5% of the full-model performance, respectively.

F.2 LLaVA-1.5-13B

Table 11 reports the results on LLaVA-1.5-13B. PCDM retains 98.4%, 97.9%, and 97.2% of the full-model performance under the 192, 128, and 64 visual token budgets, respectively. The advantage over the strongest baseline increases from 0.3% at 192 tokens to 0.9% at 64 tokens, indicating that PCDM remains robust under aggressive compression.

F.3 LLaVA-NeXT-13B

Table 12 reports the results on LLaVA-NeXT-13B. PCDM retains 97.6%, 95.4%, and 93.7% of the full-model performance under the 640, 320, and 160 visual token budgets, respectively, consistently achieving the strongest overall performance across all three budgets.

G Hyperparameter Sensitivity

We conduct controlled experiments to evaluate the sensitivity of PCDM to its main hyperparameters, including the random-walk restart probability α, the complementary content–position balancing coefficient a, the QP coefficient λ, and the diffusion weight γ. Specifically, we vary one hyperparameter at a time while fixing the remaining ones. All experiments are performed on GQA, MMBench, POPE, and TextVQA using LLaVA-1.5-7B with 64 retained visual tokens.

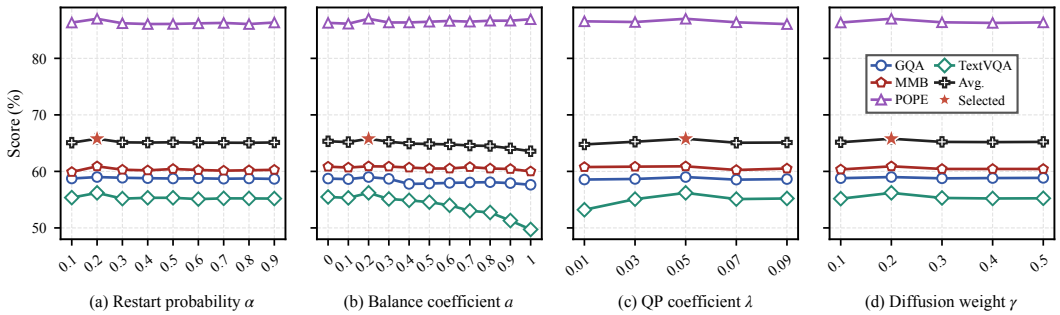
Figure 6(a) reports the results for different restart probabilities α. Performance remains stable across all four benchmarks as α varies from 0.1 to 0.9, with the highest average score achieved at $\alpha = 0.2$. This setting corresponds to a high propagation probability, enabling sufficient multi-hop information

Table 11: Performance comparison on LLaVA-1.5-13B. The 1st/2nd highest score marked by **bold** / underlines.

Method	MME	MMB	GQA	VQA^{v2}	SQA	VizWiz	VQA^{Text}	POPE	SEED^{I}	MMVet	Avg.
Upper Bound, 576 Tokens **(100%)**											
LLaVA-1.5-13B*	1829	67.8	63.2	80.0	72.7	54.3	61.2	86.0	68.2	35.7	100.0%
Retain 192 Tokens (↓66.7%)											
FastV* (ECCV'24)	**1787**	**67.3**	**61.1**	78.4	71.6	**54.8**	58.1	82.6	65.7	33.9	97.3%
CDPruner* (NeurIPS'25)	1776	66.6	60.4	78.4	72.3	52.4	58.9	**86.6**	**66.0**	34.5	97.5%
MMTok* (ICLR'26)	1769	67.2	60.2	78.4	**73.2**	54.1	59.5	86.0	65.6	35.0	98.1%
PCDM (Ours)	1782	66.9	60.1	**78.5**	72.8	53.2	**59.7**	86.2	65.7	**36.7**	**98.4%**
Retain 128 Tokens (↓77.8%)											
FastV* (ECCV'24)	**1793**	66.4	59.3	76.5	72.6	52.3	58.2	77.9	64.8	32.5	95.3%
CDPruner* (NeurIPS'25)	1778	**66.8**	59.7	77.7	73.0	52.5	58.4	**87.3**	**65.0**	35.3	97.5%
MMTok* (ICLR'26)	1766	66.2	59.2	77.7	**73.8**	**54.1**	59.0	86.0	64.5	32.8	96.9%
PCDM (Ours)	1780	66.2	**60.0**	**77.8**	73.4	53.0	**59.4**	87.0	64.9	**35.7**	**97.9%**
Retain 64 Tokens (↓88.9%)											
FastV* (ECCV'24)	1658	64.1	55.7	71.3	72.8	52.2	56.3	69.1	59.8	32.2	90.9%
CDPruner* (NeurIPS'25)	1742	65.0	59.4	**76.7**	72.5	53.2	57.9	**87.1**	64.0	34.0	96.3%
MMTok* (ICLR'26)	**1770**	**66.2**	58.5	76.0	72.7	**55.0**	58.3	84.6	63.5	31.9	95.9%
PCDM (Ours)	1745	65.4	**60.3**	76.5	**72.9**	53.4	**58.8**	86.8	**64.9**	**35.0**	**97.2%**

Table 12: Performance comparison on LLaVA-NeXT-13B. The 1st/2nd highest score marked by **bold** / underlines.

Method	MME	MMB	GQA	VQA^{v2}	SQA	VizWiz	VQA^{Text}	POPE	SEED^{I}	MMVet	Avg.
Upper Bound, 2880 Tokens **(100%)**											
LLaVA-NeXT-13B*	1885	70.2	65.4	82.9	73.4	60.4	64.2	86.3	71.9	45.5	100.0%
Retain 640 Tokens (↓77.8%)											
FastV* (ECCV'24)	**1898**	67.9	63.3	80.3	**75.6**	55.1	63.1	84.7	69.8	38.3	96.3%
CDPruner* (NeurIPS'25)	1892	68.5	63.8	**81.0**	71.6	56.5	60.7	**87.6**	70.5	39.8	96.6%
MMTok* (ICLR'26)	1864	67.5	**64.4**	76.9	73.0	**58.6**	**63.2**	85.1	66.1	41.0	96.2%
PCDM (Ours)	1890	**68.9**	63.6	80.6	73.5	56.1	62.1	87.2	**71.4**	**42.2**	**97.6%**
Retain 320 Tokens (↓88.9%)											
FastV* (ECCV'24)	1783	66.2	60.9	77.7	**75.1**	54.1	59.7	78.8	64.9	38.5	92.7%
CDPruner* (NeurIPS'25)	1807	65.9	**63.0**	**79.5**	70.8	54.5	58.8	**87.8**	68.8	41.2	94.8%
MMTok* (ICLR'26)	1840	65.7	62.2	76.7	70.1	**57.9**	**60.6**	86.7	64.2	41.5	94.6%
PCDM (Ours)	**1854**	**67.3**	62.8	79.0	70.8	54.5	59.9	87.7	**69.0**	**41.6**	**95.4%**
Retain 160 Tokens (↓94.4%)											
FastV* (ECCV'24)	1637	62.2	55.2	70.2	70.6	52.8	55.8	65.9	58.1	31.1	84.1%
CDPruner* (NeurIPS'25)	1793	65.1	62.3	**77.7**	69.9	52.8	57.5	**87.3**	67.3	38.8	92.9%
MMTok* (ICLR'26)	1800	64.9	61.7	75.6	**70.7**	**57.3**	**58.5**	85.1	63.1	**39.5**	92.9%
PCDM (Ours)	**1838**	**67.0**	**62.8**	77.4	70.1	52.0	58.4	87.1	**67.9**	39.3	**93.7%**



(a) Restart probability α (b) Balance coefficient a (c) QP coefficient λ (d) Diffusion weight γ

Figure 6: PCDM performance w.r.t. the random-walk restart probability α (a), complementary content–position balancing coefficient a (b), QP coefficient λ (c), and diffusion weight γ (d).

propagation while periodically returning the random walk to the query-conditioned initial distribution. We therefore set $\alpha = 0.2$ to balance query conditioning and relation propagation.

Figure 6(b) studies the complementary balancing coefficient a. Performance reaches its highest average score at $a = 0.2$ and remains relatively stable for small to moderate values. As a becomes larger, performance gradually decreases, with a particularly clear degradation on TextVQA. This suggests that excessively emphasizing content relevance while weakening positional relevance and semantic redundancy is less effective for text-intensive visual understanding.

Figure 6(c) compares different values of the QP coefficient λ. The results remain similar when λ varies from 0.03 to 0.09, whereas $\lambda = 0.01$ leads to a lower TextVQA score. An excessively small λ provides insufficient redundancy suppression, while increasing it beyond 0.05 brings no additional benefit. We thus set $\lambda = 0.05$ to balance relevance preservation and redundancy reduction.

Figure 6(d) shows the effect of the diffusion weight γ. The four benchmark curves remain nearly unchanged as γ increases from 0.1 to 0.5, indicating that PCDM is largely insensitive to the precise contribution of diffusion-aware redundancy. We use $\gamma = 0.2$, which already provides stable use of the propagated relation information.

Based on these observations, we set $\alpha = 0.2$, $a = 0.2$, $\lambda = 0.05$, and $\gamma = 0.2$ in all main experiments. Overall, PCDM maintains stable performance across broad hyperparameter ranges, with noticeable degradation occurring only when the complementary content–position balance approaches an extreme configuration.

H Additional Qualitative Results

We provide additional qualitative comparisons on GQA using LLaVA-1.5-7B. Figures 7 contains 6 representative examples under pruning rates of 66.7%, 77.8%, and 88.9%, corresponding to retaining 192, 128, and 64 visual tokens, respectively. Gray regions denote discarded tokens, while red rectangular boxes enclose the visual regions that appear relevant to each query based on human visual judgment. The predicted answer is shown below each image, with a green check mark indicating a correct prediction and a red cross indicating an incorrect prediction.

In these challenging examples, FastV and MMTok discard query-relevant regions and produce incorrect answers across different token budgets. In contrast, PCDM preserves the relevant objects and their surrounding context more consistently, remaining correct even under aggressive pruning.



Figure 7: Additional qualitative comparisons on GQA using LLaVA-1.5-7B. PCDM preserves query-relevant evidence more consistently than FastV and MMTok across different pruning rates.