UG-PRUNER: UNIFYING DIFFERENT VISUAL TOKEN PRUNING TASKS FOR MULTIMODAL LARGE LANGUAGE MODELS IN GRAPH

Anonymous authors
Paper under double-blind review

ABSTRACT

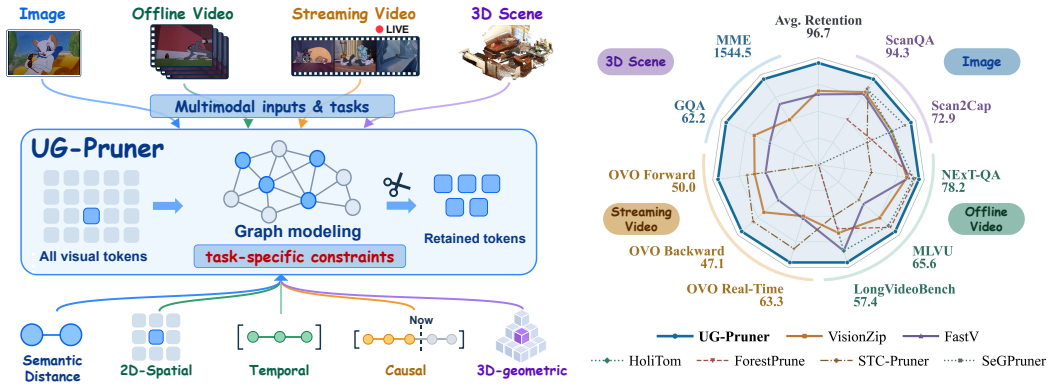
Visual token pruning is a research hot spot for reducing the computation and memory overhead of *multimodal large language models* (MLLMs). Despite great success, existing methods often require dedicated designs for different pruning tasks, such as the *image*, *video* or *3D* ones, lacking a unified and generalized paradigm. In this paper, we propose a novel and unified pruning scheme termed **UG-Pruner**. In principle, UG-Pruner formulates token pruning as a problem of graph modeling, and then unifies different pruning tasks via constrained graph building. In practice, we concretize the major properties of token pruning tasks into graph constraint terms, including the *2D spatial*, *temporal*, *3D geometry*, and *streaming causal* ones. In this case, UG-Pruner can handle different token pruning in a unified framework via configuring proper constraints to graph modeling. To validate UG-Pruner, we apply it to 4 MLLMs and conduct extensive experiments on 17 benchmarks of *image*, *off-line video*, *streaming video* and *3D scenes*. The experimental results not only show its competitive or even better performance against task-specific SOTA methods, e.g., achieving a new SOTA score of 94.27 on ScanQA, but also witness its superior generalization on all benchmarks. For instance, UG-Pruner can retain 96.72% performance under a pruning ratio up to 90% over 4 main tasks, portraying its great potential as a unified and generalized toolbox for existing MLLMs. **Our code** is given in the supplement.

1 INTRODUCTION

Recent years have witnessed the great breakthroughs made by advanced *multimodal large language models* (Li et al., 2025; Team, 2025) in the tasks of off-line or streaming video QA (Zhang et al., 2025c; Lin et al., 2024a; Di et al., 2025) and 3D scene understanding (Zheng et al., 2025; Zhu et al., 2025) in addition to the image ones (Liu et al., 2023; Li et al., 2023a). Despite great progress, the high overhead caused by the excessive visual tokens used remains a long-standing problem that plagues the practical use of MLLMs. For instance, to understand an hour-long video, Qwen3-VL-8B may use more than $350k$ visual tokens as the input, requiring more than 50GB memory overhead and minute-level latency. In this case, token pruning (or compression) (Chen et al., 2024; Yang et al., 2025b; Jiang et al., 2025; Zhang et al., 2025b; Shang et al., 2025) becomes a research hot-spot for MLLMs due to its training-free and plug-and-play merits.

The primary target of existing token pruning methods is to eliminate the token redundancy and keep the critical ones for efficient MLLMs. But in practice, practitioners often need to study specialized pruning strategies to meet the properties of different visual token pruning tasks. For instance, in 2D-image tasks, token pruning just need to estimate the redundancies or importances of visual tokens within a single image (Yang et al., 2025b; Jiang et al., 2025), while the temporal or causality factors need to be additionally considered in (streaming) video tasks (Fu et al., 2025b; Ju et al., 2026; Wang et al., 2026b). Similarly, for 3D tasks, the evaluation of visual tokens in geometry space is another important aspect for effective 3D token pruning (Li et al., 2026d). Overall, the significant diversity of token pruning tasks results in great variations in both solutions and implementations of existing methods. To this end, we question that

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.



(a) (b)

Figure 1: (a) Illustration of UG-Pruner. Via implementing proper graph constraints, UG-Pruner can unify different visual token pruning tasks in one framework. (b) Comparison with existing pruning methods. UG-Pruner not only exhibits competitive or even better performance than task-specific SOTAs, but also yields a great advantage in generalized token pruning.

"Can we establish a unified, general and easy-to-use paradigm for visual token pruning that is applicable for different MLLMs and tasks?"

The key to achieve this target lies in constructing a flexible but general theoretical framework for visual token pruning, and we notice that graph modeling yields a viable way. Firstly, applying graph modeling to process token pruning, *i.e.*, treating it as a structure feature analysis task, is a feasible solution attempted by previous works (Jiang et al., 2025; Song et al., 2026). Via effectively building a graph structure for visual tokens, their importances in an image can be estimated by using classical graph algorithms such as *information propagation* (Zhou et al., 2003). Thus, it is natural and feasible to extend graph modeling to more pruning tasks. Secondly, although different tasks exhibit different requirements for token pruning, we observe that their properties can be formulated as the concrete constraint terms for graph building. For instance, the spatial-temporal relationship (Ju et al., 2026) required for video tasks or the geometric distance (Li et al., 2026d) in 3D scenes can be concretized into the binary conditions for node connections. In this case, a flexible and unified token pruning can be achieved by incorporating task-specific constraints into graph modeling, as shown in Fig. 1a.

Motivated by the above, we propose a novel and general token pruning scheme for MLLMs, termed *Unified-Graph Pruner* (**UG-Pruner**). As illustrated in Fig. 1a, UG-Pruner defines token pruning as a task of graph modeling, which consists of the steps of *node selection*, *graph building* and *node estimation*. To enable cross-task token pruning in this framework, we also formulated the key properties of the main pruning tasks as 5 types of node connection constraints. In addition to the most common *semantic distance*, the other four constraints are *2D-spatial*, *3D-geometric*, *temporal* and *causal* ones, respectively. Then, UG-Pruner can well accomodate different task requirements. In practice, we can derive the corresponding graph structures and the final token pruning by incorporating suitable constraint terms tailored to specific tasks, thereby greatly simplifying the adaptation of token pruning across tasks. Overall, UG-Pruner offers a flexible and generalized token pruning solution via configuring proper graph constraints in a unified framework.

To validate the proposed UG-Pruner, we apply it to 4 representative MLLMs, namely LLaVa-NeXT-8B (Liu et al., 2024a), LLaVA-Video-7B (Zhang et al., 2025c), LLaVA-OneVision-7B (Li et al., 2025; Di et al., 2025) and Video-3D-LLM (Zheng et al., 2025), and conduct extensive experiments on 17 benchmarks of the image, off-line video, streaming-video and 3D scene tasks. The experimental results not only show the competitive or even better performance of UG-Pruner than existing task-specific SOTAs, *e.g.*, achiving a new SOTA performance on the benchmarks of MLVU, LongVideoBench and Scan2Cap, but also witness its generalized capability on all pruning tasks, *e.g.*, achiving a 96.72% performance retention on all benchmarks under a pruning ratio up to 90%.

Overall, our contributions are three-fold:

- We realize the first attempt of unifying different token pruning tasks for MLLMs in a flexible and general framework, such as the image, (streaming) video and 3D ones.

- We propose a novel and general token pruning scheme, termed UG-Pruner, which concretizes the main properties of different token pruning tasks into formulated constraint terms for graph modeling.
- The proposed UG-Pruner not only demonstrates competitive or even better performance than task-specific pruning methods across 17 benchmarks, but more importantly, provides an out-of-the-box solution for different MLLMs and vision-language tasks.

2 Related Work

Visual token pruning has attracted growing interest as multimodal large language models (MLLMs) expand across image, video, and 3D applications. *Image* methods select important tokens using attention or cross-modal relevance (Chen et al., 2024; Yang et al., 2025b; Cao et al., 2024; Zhang et al., 2025b; Ye et al., 2025; Song et al., 2025; Li et al., 2026a), or reduce redundancy through token merging and representative subset selection (Bolya et al., 2023; Shang et al., 2025; Zhong et al., 2025; Alvar et al., 2025; Yang et al., 2025a; Du et al., 2026). Although broadly applicable to individual frames or views, these strategies often perform suboptimally on other tasks because they overlook task-specific properties (Chen et al., 2026; Li et al., 2026b). *Offline video* methods exploit temporal relations through cross-frame selection and spatiotemporal merging (Tao et al., 2025; Fu et al., 2025b; Hyun et al., 2025; Shao et al., 2025; Li et al., 2026c; Liu et al., 2025a; Yuan et al., 2026; Wang et al., 2026a; Liu et al., 2025b), while *streaming* approaches compress incoming observations and historical states under causal access (Di et al., 2025; Zhang et al., 2025a; Wang et al., 2026b; Chen et al., 2026; Duan et al., 2026). *3D* methods consolidate repeated cross-view observations via geometry-aware grouping while preserving spatial coverage (Huang et al., 2025; Li et al., 2026d;b; Pan et al., 2026; Jiang et al., 2026). Several token pruning methods have also explored graph-based modeling. G-Prune (Jiang et al., 2025) propagates information over semantic token graphs, Script (Yang et al., 2025c) combines graph-structured pruning with query relevance, and ForestPrune (Ju et al., 2026) models video tokens as spatiotemporal forests. These methods apply graph modeling to specific input settings, however, lacking a generalized definition and a feasible scheme for unified token pruning.

3 Method

3.1 Preliminaries

The primary target of visual token pruning (or compression) (Chen et al., 2024; Jiang et al., 2025) is to reduce the number of input visual tokens while retaining the performance of MLLMs.

Specifically, given a visual input denoted as V, which can be an image, a video or multi-view 3D images, and a query Q, an MLLM uses a vision encoder to extract its visual token features, denoted as $\mathbf{F}_v \in \mathbb{R}^{N\times d}$, where N denotes the number of visual tokens and d is the feature dimension. Then, the objective of token pruning can be defined as

$$\mathcal{P}^\star = \arg\min_{\mathcal{P}} \mathcal{D}(\mathcal{M}(\mathbf{F}_v, Q), \mathcal{M}(\mathbf{F}'_v, Q)) \,. \tag{1}$$

Here, $\mathcal{M}(\cdot)$ denotes the MLLM, and $\mathbf{F}'_v$ are the retained token features after pruning with $|\mathbf{F}'_v| = K < N$ for a predefined token budget K, $\mathcal{D}(\cdot,\cdot)$ measures the difference between predictions before and after pruning. $\mathcal{P}^\star$ is the optimal pruning strategy solved, which is also categorized into the types of *out-LLM* and *in-LLM* (Zhang et al., 2025b), respectively. In this paper, we mainly focus on the *out-LLM* pruning strategy, *i.e.*, pruning tokens before the LLM.

In terms of the pruning solutions, existing methods vary greatly across different VL tasks. As previously mentioned, compared to image-based pruning, token pruning for videos requires addressing token redundancy across spatiotemporal dimensions (Fu et al., 2025b; Yuan et al., 2026), whereas 3D perception tasks necessitate the evaluation of token importance within geometric space (Li et al., 2026d;b). Given this context, it is difficult to construct a unified toolbox that facilitates convenient token pruning for current MLLMs on different tasks in a shared framework.

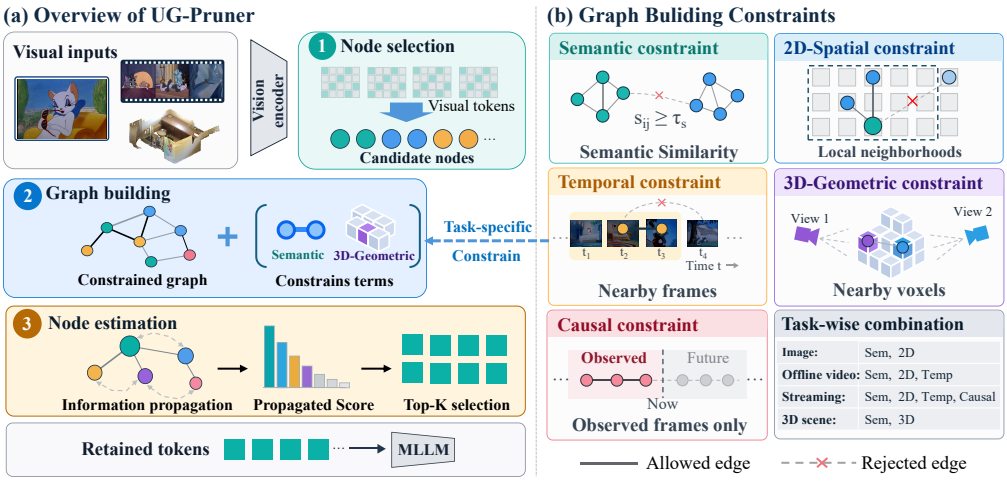


Figure 2: Overview of UG-Pruner. UG-Pruner presents a unified framework for different visual token pruning (a), which consists of the steps of *node selection*, *graph building* and *node estimation*. To generalize across tasks, UG-Pruner formulates 5 main properties of existing pruning tasks into graph constraint terms (b). Via implementing suitable constraints, UG-Pruner can construct effective token graphs for different pruning tasks.

3.2 Method Overview

In this paper, we propose a novel and unified pruning scheme for existing MLLMs, termed **UG-Pruner**, of which framework is illustrated in Fig. 2. UG-Pruner formulates the specific properties of different vision-language tasks as mathematical constraints for graph construction, enabling task-wise important tokens to be selected through a shared graph-based pruning framework.

Specifically, given an image, or alternatively, a video or multi-view 3D images, the primary step of UG-Pruner is to build a sparse and weighted graph for visual tokens, denoted by $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, W\}$. Here, $\mathcal{V}$ is a set of graph nodes represented by visual tokens, and $\mathcal{E}$ denotes the the graph edges with W as the edge weight matrix. To reduce the cost of graph modeling, we will first select M candidates from all visual tokens via image-wise random selection or existing image pruning methods.

In practice, we represent $\mathcal{E}$ by a binary matrix $E \in \mathbb{R}^{M\times M}$. And the edges are regularized by a set of task-specific constraints, denoted by $\mathcal{C}$. An edge is constructed only when the node pair $r_{i,j}$ satisfies the task-specific constraints, defined by

$$E_{ij} = \begin{cases} 1, & \text{if } r_{i,j} \text{ meets } \mathcal{C}_{task}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

then the weighted adjacency matrix W are valued by node-pair cosine similarities:

$$W_{ij} = E_{ij} \frac{f_i^\top f_j}{\|f_i\|_2 \|f_j\|_2}, \tag{3}$$

where $f_i \in \mathbf{F}_v$ denotes the token feature of node v_i. In this way, we can obtain a sparse and weighted graph to represent the structural relationships between visual tokens.

Node estimation. Given the constructed token graph $\mathcal{G}$, we then apply a classical graph algorithm to evaluate node importance. Similar to (Jiang et al., 2025), we use *information propagation* in our paper. Specifically, the score of each node is initialized by its token feature norm, *i.e.*, $S_i^{(0)} = \|\mathbf{f}_i\|_2$. Let $\widehat{W}$ be the row-normalized form of W, the scores at the t-th step are updated by

$$\mathbf{S}^{(t+1)} = \widehat{W}^\top \mathbf{S}^{(t)}, \qquad \text{where } t = 0, \ldots, T-1. \tag{4}$$

The propagated score $S_i^{(T)}$ aggregates information from semantically related nodes that satisfy the task structure. After propagation, UG-Pruner select the top-K highest scored nodes as the preserved

visual tokens given the predefined pruning target. In practice, UG-Pruner will empirically define suitable constraint terms $\mathcal{C}$ to regularize the graph building, *i.e.*, Eq. 2, thereby obtaining proper structure information for different pruning tasks.

3.3 Graph Building Constraints

We summarize five main properties of existing token pruning tasks (Jiang et al., 2025; Ju et al., 2026; Li et al., 2026d), and concretize them into graph constraint terms, including the semantic ($Sem.$), 2D spatial ($2D$), temporal ($Temp.$), causal ($Causal.$) and 3D geometric ($3D$) ones.

Semantic constraint (C_{Sem}). Semantic constraint provides a basic criterion for node connection in UG-Pruner. In UG-Pruner, we use feature similarity-based thresholding for this constraint. Thus the edge definition in Eq. 2 becomes:

$$E_{ij}^{Sem} = \begin{cases} 1, & \text{if } s_{ij} \geq \tau_s, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Here s_{ij} is the cosine similarity between node pairs, and $\tau_s \in (0, 1)$ is the similarity threshold.

2D spatial constraint (C_{2D}). This constraint is to preserves the local structure of an 2D image, better facilitating the token coverage (Du et al., 2026). Its definition is

$$E_{ij}^{2D} = \begin{cases} 1, & \text{if } \|x_i - x_j\|_2 \leq \rho_s, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

Here $x_i \in \mathbb{R}^2$ denotes the coordinates of token i, and ρ_s is the spatial neighborhood radius.

Temporal constraint (C_{Temp}). Temporal redundancy is a key issue for video token pruning (Ju et al., 2026). Thus, this constraint is used to regularize graph building from the temporal aspect:

$$E_{ij}^{Temp} = \begin{cases} 1, & \text{if } |t_i - t_j| \leq \Delta_t, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Here, t_i denotes the frame index of token i, and Δ_t is the temporal neighborhood radius.

Causal constraint (C_{Causal}). Causality is a key property of tasks like streaming videos (Chen et al., 2026; **?**), *i.e.*, upcoming frames are unknown for token pruning. We formulate this property as a *Causal* constraint for UG-Pruner, which is defined by

$$E_{ij}^{Causal} = \begin{cases} 1, & \text{if } \max(t_i, t_j) \leq h, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Here h denotes the latest observed frame index.

3D geometric constraint (C_{3D}). Geometric proximity is important for associating tokens that observe nearby regions across different views of a 3D scene (Li et al., 2026d). Thus, this constraint is used to establish cross-view connections in the shared 3D space. It is defined by

$$E_{ij}^{3D} = \begin{cases} 1, & \text{if } \nu_i \neq \nu_j \text{ and } d_{ij} \leq r, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Here ν_i denotes the view index of token i, $d_{ij} = \|\mathbf{z}_i - \mathbf{z}_j\|_\infty$ is the *voxel* distance (Liu et al., 2019) between two tokens in the shared 3D space, and r is the voxel neighborhood radius.

In practice, UG-Pruner allows for the flexible combination of the above constraint terms to achieve an optimal yet task-specific pruning in a unified framework.

3.4 Specific Graph Constraints for Different Tasks

UG-Pruner can flexibly combine the constraint terms for task-specific graph modeling in a unified graph modeling framework. The task-wise settings are given below.

Image tasks. For image tasks, we mainly consider to reduce the semantic redundancy while keeping a proper coverage in 2D space. Thus, the graph constraints used are C_{Sem} and C_{2D}.

Table 1: The constraint configurations of UG-Pruner for different pruning tasks.

Tasks	Settings	Tasks	Settings
Image	$C_{\text{sem}} \vee C_{\text{2D}}$	Offline video	$C_{\text{sem}} \wedge C_{\text{2D}} \wedge C_{\text{temp}}$
3D scene	$C_{\text{sem}} \vee C_{\text{3D}}$	Streaming video	$C_{\text{sem}} \wedge C_{\text{2D}} \wedge C_{\text{temp}} \wedge C_{\text{causal}}$

Offline video tasks. Building upon the settings of image pruning, we further introduce the temporal constraints to off-line video understanding tasks. It mainlys aim is to reduce the token redundancy across frames. Thus, the constraints for off-line video are C_{Sem}, C_{2D} and C_{Temp}.

Streaming video tasks. Compared with the off-line video tasks, the streaming ones have strict causality, *i.e.*, the visual tokens are obtained frame-by-frame rather than from a complete video. Thus, we add the *Causal* constraint in addition to the off-line video ones. In practice, we will construct the graph frame by frame and perform node estimation after each short interval.

3D understanding tasks. Similar to the constraints of image tasks, we extend 3D tasks with the *geometry* constraint term, thereby reducing the token redundancy in the geometrty space. In practice, we adopt the widely-used multi-view image setting (Li et al., 2026b) for 3D scene representation.

In addition to the above configuration of constraint terms, the implementation of UG-Pruner to different type of MLLMs are also introduced in Sec. 4.1 and detailed in Appendix.

4 Experiments

4.1 Experimental Setup

Models and Baselines. We evaluate UG-Pruner on four representative MLLMs covering different tasks, including LLaVA-NeXT-8B (Liu et al., 2024a) for images, LLaVA-Video-7B (Zhang et al., 2025c) for offline videos, LLaVA-OneVision-7B (Li et al., 2025) for streaming videos, and Video-3D-LLM (Zheng et al., 2025) for multi-view 3D scenes. In terms of comparison, the image-specific pruning methods are FastV (Chen et al., 2024), VisionZip (Yang et al., 2025b), DivPrune (Alvar et al., 2025), and CoIN (Du et al., 2026), the off-line video approaches inculde HoliTom (Shao et al., 2025), STTM (Hyun et al., 2025), FrameFusion (Fu et al., 2025b), and ForestPrune (Ju et al., 2026), the streaming pruning methods include VidCom^2 (Liu et al., 2025a), and STC-Pruner (Wang et al., 2026b), and the 3D pruning methods are SeGPruner (Li et al., 2026d) and Geo3DPruner (Li et al., 2026b). For a more comprehensive comparison, we apply several image pruning approaches to all benchmarks since they can be essentially regarded as image-based pruning tasks. Besides, the multi-view 3D pruning is also considered as a multi-frame based setting, where the video methods can be validated for comparison. To ensure fairness, we do not alter their default designs of these transferred methods. More detail can refer to our Appendix.

Datasets and Metrics. We evaluate image understanding on GQA (Hudson & Manning, 2019), MMBench (Liu et al., 2024b), POPE (Li et al., 2023b), MME (Fu et al., 2023), and TextVQA (Singh et al., 2019). Video evaluation covers NExT-QA (Xiao et al., 2021), Video-MME (Fu et al., 2025a), MLVU (Zhou et al., 2025), LongVideoBench (Wu et al., 2024), and MVBench (Li et al., 2024) for offline understanding, along with OVO-Bench (Niu et al., 2025) and StreamingBench (Lin et al., 2024b) for streaming understanding. For 3D scenes, we use ScanRefer (Chen et al., 2020), Multi3DRefer (Zhang et al., 2023), Scan2Cap (Chen et al., 2021), ScanQA (Azuma et al., 2022), and SQA3D (Ma et al., 2023). We follow the standard evaluation metrics of each benchmark. Image and video understanding are evaluated using their official accuracy or aggregate scores. Detailed evaluation protocols are provided in Appendix A.3.

Implementation Details. We perform five graph propagation steps for all settings. The semantic threshold is set to $\tau_s = 0.5$ for image, offline-video, and 3D tasks, and $\tau_s = 0.8$ for streaming video. For the task-specific constraints summarized in Table 1, we set the 2D spatial radius to $\rho_s = 6$, the offline temporal window to $\Delta_t = 5$, and the 3D voxel radius to $r = 1$. Following prior streaming-video evaluations (Wang et al., 2026b), we sample videos at 0.5 FPS, and the streaming compression window is set to 16. Unless otherwise specified, the pruning ratio is set to 90% for image, offline video, and 3D tasks, and 75% for streaming video. Latency is measured on a single NVIDIA A800 GPU. More details refer to our Appendix.

Table 2: Comparison with task-specific methods. *Ratio* denotes the pruning rate. ‡ uses a tuned 3D-MLLM that has not been released, so we only report its performance retention. The best and second-best results are in **bold** and underlined, respectively.

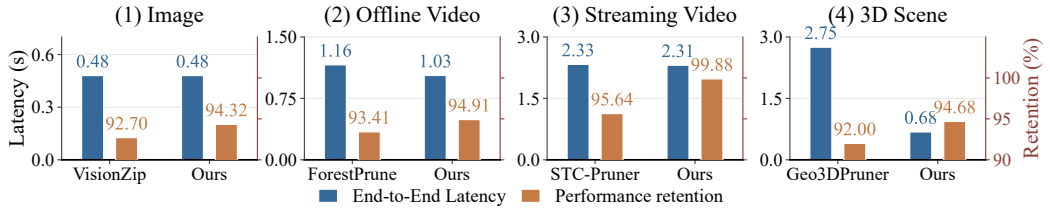
Images — LLaVA-NeXT-8B							
Method	Ratio	GQA	MMBench	POPE	MME	TextVQA	Avg. Ret.
Baseline	0%	65.38	72.08	87.84	1587.47	65.41	100.00
FastV (Chen et al., 2024)	90%	60.20	**68.30**	83.01	1488.01	53.53	91.38
VisionZip (Yang et al., 2025b)	90%	60.91	67.10	83.21	1452.66	**59.53**	92.70
DivPrune (Alvar et al., 2025)	90%	**63.32**	67.01	86.34	1507.54	53.73	93.04
CoIN (Du et al., 2026)	90%	63.09	66.92	**87.28**	1535.03	56.29	94.29
UG-Pruner	90%	62.18	68.12	85.36	**1544.45**	57.26	**94.32**
Offline Video Understanding — LLaVA-Video-7B							
Method	Ratio	NExT-QA	VideoMME	MLVU	LongVideoBench	MVBench	Avg. Ret.
Baseline (Zhang et al., 2025c)	0%	80.97	63.96	69.04	60.14	61.80	100.00
FastV (Chen et al., 2024)	80%	78.85	61.26	62.13	57.07	59.23	94.83
VisionZip (Yang et al., 2025b)	80%	79.59	61.30	65.77	56.84	60.68	96.51
HoliTom (Shao et al., 2025)	80%	79.55	**62.59**	66.30	57.29	**61.13**	**97.31**
STTM (Hyun et al., 2025)	80%	**80.90**	59.80	67.10	55.90	59.80	95.45
FrameFusion (Fu et al., 2025b)	80%	78.92	60.63	66.12	**57.97**	59.92	96.32
ForestPrune (Ju et al., 2026)	80%	79.68	62.30	66.76	56.99	60.05	96.98
UG-Pruner	80%	79.18	60.85	**67.30**	57.44	60.77	96.91
FastV (Chen et al., 2024)	90%	77.21	58.48	59.26	55.72	57.63	91.78
VisionZip (Yang et al., 2025b)	90%	77.00	59.30	62.53	53.18	58.10	92.32
HoliTom (Shao et al., 2025)	90%	77.66	**61.59**	64.28	55.72	**59.80**	**94.98**
STTM (Hyun et al., 2025)	90%	65.00	56.00	64.80	54.50	56.00	88.05
FrameFusion (Fu et al., 2025b)	90%	77.24	57.81	58.58	55.05	57.12	91.04
ForestPrune (Ju et al., 2026)	90%	77.83	60.67	64.70	52.58	57.98	93.41
UG-Pruner	90%	**78.20**	59.19	**65.64**	**57.37**	58.40	94.91
Streaming Video Understanding — LLaVA-OneVision-7B, 0.5 FPS							
Method	Ratio	OVO-RT	OVO-Backward	OVO-Forward	OVO Avg.	StreamingBench	Avg. Ret.
Baseline (Di et al., 2025)	0%	64.40	46.70	47.80	52.97	72.46	100.00
VisionZip (Yang et al., 2025b)	75%	53.80	44.00	44.70	47.50	70.22	93.29
VidCom^2 (Liu et al., 2025a)	75%	60.40	45.60	45.20	50.40	69.30	95.40
STC-Pruner (Wang et al., 2026b)	75%	60.50	45.50	45.80	50.60	69.38	95.64
UG-Pruner	75%	**63.32**	**47.09**	**50.05**	**53.49**	**71.58**	**99.88**
3D Scene Understanding — Video-3D-LLM							
Method	Ratio	ScanRefer IoU@.25/.5	Multi3DRefer IoU@.25/.5	Scan2Cap B-4/CIDEr	ScanQA CIDEr/EM	SQA3D EM	Avg. Ret.
Baseline (Zheng et al., 2025)	0%	58.16/51.77	57.35/52.10	40.56/80.28	101.28/30.01	58.43	100.00
FastV (Chen et al., 2024)	80%	56.26/49.94	55.58/50.47	35.60/67.71	95.80/28.26	54.96	93.55
CoIN (Du et al., 2026)	80%	56.17/49.87	55.72/50.60	37.80/71.94	96.38/28.43	56.69	95.21
SeGPruner (Li et al., 2026d)	80%	55.65/49.48	53.89/49.09	**39.38/78.43**	95.07/28.34	56.15	95.40
Geo3DPruner‡ (Li et al., 2026b)	80%	–	–	–	–	–	96.30
UG-Pruner	80%	**56.64/50.37**	**55.96/50.89**	39.23/75.54	**96.65/28.83**	**56.95**	**96.64**
FastV (Chen et al., 2024)	90%	55.67/49.38	53.01/48.17	33.27/59.46	91.02/26.95	54.65	89.48
CoIN (Du et al., 2026)	90%	**55.81/49.61**	54.31/49.38	35.22/60.33	91.76/27.20	55.38	91.03
SeGPruner (Li et al., 2026d)	90%	54.74/48.67	52.25/47.56	36.89/69.17	92.11/27.47	55.92	91.76
Geo3DPruner‡ (Li et al., 2026b)	90%	–	–	–	–	–	92.00
UG-Pruner	90%	55.63/49.34	**54.88/49.86**	**38.86/72.90**	**94.27/28.04**	**56.46**	**94.68**

4.2 Quantitative Analysis

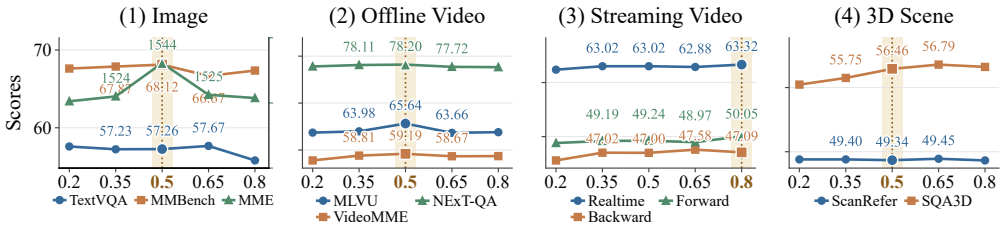
Task-specific Comparisons. We first compare UG-Pruner with task-specific pruning methods on the benchmarks of the image, offline-video, streaming-video and 3D understanding tasks in Table 2. From this table, we can first observe that task-specific methods achieve strong performance on their target benchmarks. For instance, CoIN and HoliTom retain 94.29% and 94.98% performance on image and offline-video benchmarks, while STC-Pruner and SeGPruner achieve 95.64% and 91.76% retention on streaming-video and 3D benchmarks, respectively. These results indicate that effective token pruning requires appropriate modeling of task-specific properties. Nevertheless, it can also be seen that UG-Pruner yields competitive or better results than these task-specific methods. Notably, UG-Pruner achieves new SOTA performance on multiple benchmarks, *e.g.*, +2.89 points on OVO-Bench and +2.16 points on ScanQA. Overall, these superior results well confirm the effectiveness of graph modeling as well as the summarized constraints in UG-Pruner for existing pruning tasks.

Table 3: Comparison across tasks. Methods marked with ○, △, ◇, and □ are originally designed for image, offline-video, streaming-video, and 3D-scene pruning under a pruning ratio up to 90%, respectively. *Avg.* is the mean baseline-normalized performance retention over reported metrics.

	Images		Offline Video			Streaming Video			3D Scenes		
Method	GQA	MME	NExT-QA	LongVideoBench	MLVU	Real-Time	Backward	Forward	ScanQA	Scan2Cap	Avg.
Baseline	65.38	1587.47	80.97	60.14	69.04	64.40	46.70	47.80	101.28	80.28	100.00
FastV○	60.20	1488.01	77.21	55.72	59.26	54.24	42.00	43.10	91.02	59.46	88.79
VisionZip○	60.91	1452.66	77.00	53.18	62.53	53.80	44.00	44.70	91.44	61.22	89.66
HoliTom△	–	–	77.66	55.72	64.28	–	–	–	92.17	60.97	89.72
ForestPrune△	–	–	77.83	52.58	64.70	–	–	–	85.39	49.01	84.52
STC-Pruner◇	–	–	73.52	50.79	58.33	60.50	45.50	45.80	92.42	45.89	86.92
SeGPruner□	–	–	–	–	–	–	–	–	92.11	69.17	88.55
UG-Pruner	**62.18**	**1544.45**	**78.20**	**57.37**	**65.64**	**63.32**	**47.09**	**50.05**	**94.27**	**72.90**	**96.72**



(a) Efficiency comparison with task-specific pruning methods.



(b) Semantic threshold sensitivity.

Figure 3: The efficiency comparison (a) between UG-Pruner and existing methods, and the sensitivity tuning (b) of the semantic constraint of UG-Pruner on four main tasks.

Cross-task generalization comparison. We further compare UG-Pruner with representative pruning methods across multiple tasks in Table 3. From this table, we can first observe that despite generalization, image-based methods often perform poorly on other tasks. For instance, VisionZip lags behind the HoliTom and STC-Pruner by 1.75 and 6.70 points on MLVU and OVO-RT, respectively. Meanwhile, streaming pruning methods also exhibit a certain flexibility, which can accommodate both off-line video and 3D pruning tasks. However, they also encounter drastic drops when being generalized. These results suggest that task-specific properties are still the key factors affecting the quality of token pruning. In stark contrast, our UG-Pruner exhibit the best generalized performance across all tasks. Compared with the image pruning methods, UG-Pruner achieve obvious average performance gains on 4 tasks, *e.g.*, +7.06% *avg.* performance than VisionZip. Overall, these results well confirm the unified and generalized token pruning achieved by our UG-Pruner.

Efficiency Comparison. We further compare the efficiency of UG-Pruner with representative pruning methods in Fig. 3a. We can first see that while offering superior generalization, UG-Pruner does not incur higher computation or memory overhead than task-specific methods.In contrast, it show better efficiency than some task-specific methods, *e.g.*, reducing 75.3% latency than Geo3DPruner on ScanQA. These results further confirm our merits as a unified tool-box for existing MLLMs.

Ablation Studies. We ablate the main designs of our UG-Pruner in Tab. 4. The first block shows the impact of node selection. From these results, we can see that different node selection strategies have a certain but limited influence on the subsequent graph modeling. For instance, the variation in average performance across three node selection schemes remains within 1.3%.

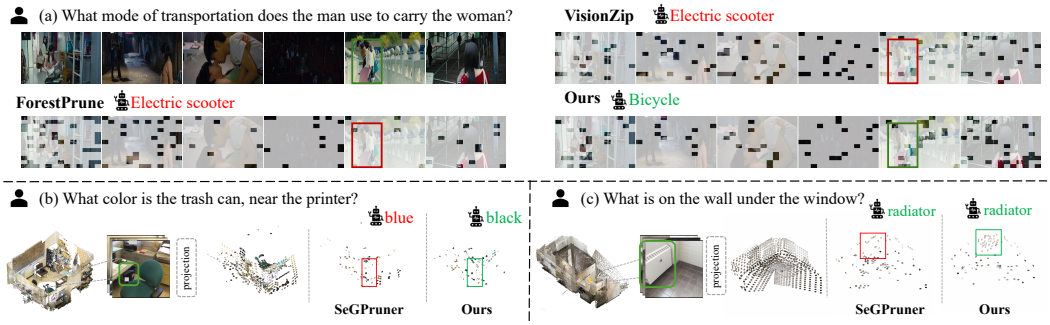


Figure 4: Visualizations of our UG-Pruner and the compared methods. The green boxes denote the task-relevant tokens that are correctly retained, while the red ones indicate mistakenly pruned ones. Retaining high generality, our UG-Pruner can compete with task-specific pruning methods, primarily due the effective graph modeling based on appropriate constraints configured.

To explain, node selection in UG-Pruner only aims to quickly identify a small set of candidate nodes, while the substantial redundancy among visual tokens makes this coarse selection relatively easy.

In the second block of Tab. 4, we examine the benefit of using task-specific constraints in addition to the semantic one. It can be clearly observed that these constraints are critical to UG-Pruner. Using the semantic constraint alone leads to a notable performance drop, *e.g.*, -3.67 points on LongVideoBench. In addition, we further analyze the influence of the semantic constraint C_{Sem} across different pruning tasks in Fig. 3b. From these plots, we can first see that most tasks exhibit similar requirements for semantic connections, where a threshold of 0.5 provides consistently stable performance. The only exception is streaming video, which favors a higher semantic threshold. This difference can be attributed to its *causal* setting, where UG-Pruner estimates token importance incrementally as each frame arrives, making semantic similarity a more direct basis for graph construction. Overall, these results show that semantic similarity serves as an important and general constraint, while additional constraints should be introduced according to the properties of each task.

Table 4: Ablation of the candidate selector and graph affinity. ♯ marks the final configuration.

Settings	LongVB	ScanQA	StreamingBench	Avg.
Node Selection				
Random	57.52	92.87	70.54	94.90
VisionZip	57.22	**94.33**	71.18	95.51
G-Prune♯	**57.37**	94.27	**71.58**	**95.75**
Affinity Design				
Semantic	53.70	91.75	69.98	92.15
Constraints♯	**57.37**	**94.27**	**71.58**	**95.75**

4.3 Qualitative Analysis

We can first see that in Fig. 4-(a), ForestPrune concentrates the retained tokens in the earlier frames, whereas our spatio-temporal constraints reduce inter-frame redundancy while maintaining a more balanced temporal distribution. For 3D scenes, the geometric constraints preserve salient regions with broader spatial coverage, for example, UG-Pruner retains the trash-can region in Fig. 4-(b) and produces the correct answer. Overall, these results further validate the effectiveness of adapting graph constraints to different input structures.

5 Conclusion

In this paper, we introduce UG-Pruner, a novel and unified training-free token pruning framework for efficient MLLMs across tasks. UG-Pruner formulates visual token pruning as constrained graph modeling, where the task-specific properties are transformed into node-connection constraints. In this case, UG-Pruner can unify different token pruning tasks in a general framework via configuring proper constraint terms. The extensive experiments on 4 MLLMs and 17 benchmarks of 4 pruning tasks not only demonstrate the competitive or even superior performance of UG-Pruner against task-specific SOTA methods, but also validate its strong generalization as a unified pruning framework across diverse MLLMs and token pruning tasks.

AI USE STATEMENT

We used generative AI tools only for organizing and restructuring existing code during the research process. Generative AI tools were not used for research ideation, methodology development, experimental design, data analysis. All research ideas, technical contributions, experimental results, and the final manuscript were developed and prepared by the authors. We have reviewed all AI-assisted outputs and take full responsibility for the final content of this work.

REFERENCES

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. DivPrune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9392–9401, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Alvar_DivPrune_Diversity-based_Visual_Token_Pruning_for_Large_Multimodal_Models_CVPR_2025_paper.html.

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. URL https://openaccess.thecvf.com/content/CVPR2022/html/Azuma_ScanQA_3D_Question_Answering_for_Spatial_Scene_Understanding_CVPR_2022_paper.html.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=JroZRaRw7Eu.

Jianjian Cao, Peng Ye, Shengze Li, Chong Yu, Yansong Tang, Jiwen Lu, and Tao Chen. MADTP: Multimodal alignment-guided dynamic token pruning for accelerating vision-language transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15710–15719, 2024.

Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3d object localization in RGB-D scans using natural language. In *European Conference on Computer Vision*, pp. 202–221. Springer, 2020.

Dave Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X. Chang. Scan2Cap: Context-aware dense captioning in RGB-D scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3193–3203, 2021. doi: 10.1109/CVPR46437.2021.00321.

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. doi: 10.1007/978-3-031-73004-7_2.

Xueyi Chen, Keda Tao, Kele Shao, and Huan Wang. Streamingtom: Streaming token compression for efficient video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24675–24685, June 2026.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Tao Zhong, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video KV-cache retrieval. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=8g9fs6mdEG.

Chenxi Du, Yongheng Deng, Jiani Liu, Yujia Zhang, Xi Chen, and Ju Ren. Coin: Coverage and informativeness-guided token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10492–10501, June 2026.

Moxu Duan, Jingwen Fu, and Yuwang Wang. Think in sets for streaming video token compression. *arXiv preprint arXiv:2608.01169*, 2026. doi: 10.48550/arXiv.2608.01169. URL https://arxiv.org/abs/2608.01169.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023. doi: 10.48550/arXiv.2306.13394.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025a. doi: 10.1109/CVPR52734.2025.02245.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22654–22663, 2025b.

Hsiang-Wei Huang, Fu-Chen Chen, Wenhao Chai, Che-Chun Su, Lu Xia, Sanghun Jung, Cheng-Yen Yang, Jenq-Neng Hwang, Min Sun, and Cheng-Hao Kuo. Zero-shot 3d question answering via voxel-based dynamic token compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19424–19434, 2025.

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. doi: 10.1109/CVPR.2019.00686.

Jeongseok Hyun, Sukjun Hwang, Su Ho Han, Taeoh Kim, Inwoong Lee, Dongyoon Wee, Joon-Young Lee, Seon Joo Kim, and Minho Shim. Multi-granular spatio-temporal token merging for training-free acceleration of video llms. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 23990–24000. IEEE, 2025. doi: 10.1109/ICCV51701.2025.02224. URL https://doi.org/10.1109/ICCV51701.2025.02224.

Jerry Jiang, Haowen Sun, Denis Gudovskiy, Yohei Nakata, Tomoyuki Okuno, Kurt Keutzer, and Wenzhao Zheng. Proxy3d: Efficient 3d representations for vision-language models via semantic clustering and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23816–23825, June 2026.

Yutao Jiang, Qiong Wu, Wenhao Lin, Wei Yu, and Yiyi Zhou. What kind of visual tokens do we need? training-free visual token pruning for multi-modal large language models from the perspective of graph. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 4075–4083. AAAI Press, 2025. doi: 10.1609/AAAI.V39I4.32427. URL https://doi.org/10.1609/aaai.v39i4.32427.

Shaobo Ju, Baiyang Song, Tao Chen, Jiapeng Zhang, Qiong Wu, Chao Chang, Huaixi Wang, Yiyi Zhou, and Rongrong Ji. Forestprune: High-ratio visual token compression for video multimodal large language models via spatial-temporal forest modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8326–8336, June 2026.

Ao Li, Yuxiang Duan, Jinghui Zhang, Congbo Ma, Yutong Xie, Gustavo Carneiro, Mohammad Yaqub, and Hu Wang. TransPrune: Token transition pruning for efficient large vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39529–39538, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_TransPrune_Token_Transition_Pruning_for_Efficient_Large_Vision-Language_Model_CVPR_2026_paper.html.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=zKv8qULV6n.

Han Li, Zehao Huang, Jiahui Fu, Naiyan Wang, and Si Liu. Geometry-guided 3d visual token pruning for video-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9615–9625, June 2026b.

Jinlong Li, Liyuan Jiang, Haonan Zhang, and Nicu Sebe. Token reduction via local and global contexts optimization for efficient video large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10451–10461, June 2026c.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023a. URL https://proceedings.mlr.press/v202/li23q.html.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_MVBench_A_Comprehensive_Multi-modal_Video_Understanding_Benchmark_CVPR_2024_paper.html.

Wenli Li, Kai Zhao, Haoran Jiang, Enquan Yang, Yi Su, and Dan Zeng. SeGPruner: Semantic-geometric visual token pruner for 3d question answering. *arXiv preprint arXiv:2603.29437*, 2026d. doi: 10.48550/arXiv.2603.29437.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne 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. Association for Computational Linguistics, 2023b. doi: 10.18653/v1/2023.emnlp-main.20.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984, 2024a. doi: 10.18653/v1/2024.emnlp-main.342.

Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. StreamingBench: Assessing the gap for MLLMs to achieve streaming video understanding. *CoRR*, abs/2411.03628, 2024b. doi: 10.48550/arXiv.2411.03628.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

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 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Xuyang Liu, Yiyu Wang, Junpeng Ma, and Linfeng Zhang. Video compression commander: Plug-and-play inference acceleration for video large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1910–1924, 2025a. doi: 10.18653/v1/2025.emnlp-main.98.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pp. 216–233. Springer, 2024b. doi: 10.1007/978-3-031-72658-3_13.

Yudong Liu, Jingwei Sun, Yueqian Lin, Jianyi Zhang, Jingyang Zhang, Ming Yin, Qinsi Wang, Hai Li, and Yiran Chen. Keyframe-oriented vision token pruning: Enhancing efficiency of large vision language models on long-form video processing. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pp. 20802–20811. IEEE, 2025b. doi: 10.1109/ICCV51701.2025.01934. URL https://doi.org/10.1109/ICCV51701.2025.01934.

Zhijian Liu, Haotian Tang, Yujun Lin, and Song Han. Point-voxel CNN for efficient 3d deep learning. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 963–973, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/5737034557ef5b8c02c0e46513b98f90-Abstract.html.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3D: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=IDJx97BC38.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, and Jiaqi Wang. OVO-Bench: How far is your video-LLMs from real-world online video understanding? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18902–18913, 2025. doi: 10.1109/CVPR52734.2025.01761.

Tianbo Pan, Xingyi Yang, and Xinchao Wang. Merge3d: Efficient 3d multimodal llms via joint 2d-3d token merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 31066–31077, June 2026.

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

Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. HoliTom: Holistic token merging for fast video large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4525.

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. doi: 10.1109/CVPR.2019.00851.

Dingjie Song, Wenjun Wang, Shunian Chen, Xidong Wang, Michael X. Guan, and Benyou Wang. Less is more: A simple yet effective token reduction method for efficient multi-modal LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 7614–7623. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.508/.

Jouwon Song, Sohyeon Kim, and Kyeongbo Kong. Uncertainty-guided graph formulation via MWIS for token pruning in LVLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 9510–9519, 2026. URL https://openaccess.thecvf.com/content/CVPR2026F/html/Song_Uncertainty-Guided_Graph_Formulation_via_MWIS_for_Token_Pruning_in_LVLMs_CVPRF_2026_paper.html.

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. DyCoke: Dynamic compression of tokens for fast video large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18992–19001, 2025.

Qwen Team. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025. doi: 10.48550/ARXIV.2511.21631. URL https://doi.org/10.48550/arXiv.2511.21631.

Hesong Wang, Xin Jin, Lu Lu, Chenhaowen Li, Jian Chen, Qiang Liu, and Huan Wang. Earlytom: Early token compression completes fast video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40559–40568, June 2026a.

Yiyu Wang, Xuyang Liu, Xiyan Gui, Xinying Lin, Boxue Yang, Chenfei Liao, Tailai Chen, and Linfeng Zhang. Accelerating streaming video large language models via hierarchical token compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18523–18533, June 2026b.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/329ad516cf7a6ac306f29882e9c77558-Abstract-Datasets_and_Benchmarks_Track.html.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9777–9786, 2021.

Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Chendi Li, Jinghua Yan, Yu Bai, Ponnuswamy Sadayappan, Xia Hu, and Bo Yuan. TopV: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19803–19813, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_TopV_Compatible_Token_Pruning_with_Inference_Time_Optimization_for_Fast_CVPR_2025_paper.html.

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, 2025b.

Zhongyu Yang, Dannong Xu, Wei Pang, and Yingfang Yuan. Script: Graph-structured and query-conditioned semantic token pruning for multimodal large language models. *Trans. Mach. Learn. Res.*, 2025, 2025c. URL https://openreview.net/forum?id=F6xKzbgcHq.

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. ATP-LLaVA: Adaptive token pruning for large vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24972–24982, 2025. doi: 10.1109/CVPR52734.2025.02325.

Chao Yuan, Shimin Chen, Minliang Lin, Limeng Qiao, Guanglu Wan, and Lin Ma. Unicomp: Rethinking video compression through informational uniqueness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18609–18618, June 2026.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-VStream: Efficient real-time understanding for long video streams. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21059–21069, 2025a.

Yiming Zhang, ZeMing Gong, and Angel X. Chang. Multi3DRefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15225–15236, 2023. doi: 10.1109/ICCV51070.2023.01397. URL https://openaccess.thecvf.com/content/ICCV2023/html/Zhang_Multi3DRefer_Grounding_Text_Description_to_Multiple_3D_Objects_ICCV_2023_paper.html.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A. Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. SparseVLM:

Visual token sparsification for efficient vision-language model inference. In *Forty-Second International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*. PMLR, 2025b. URL https://proceedings.mlr.press/v267/zhang25s.html.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-Video: Video instruction tuning with synthetic data. *Transactions on Machine Learning Research*, 2025c. URL https://openreview.net/forum?id=EE1FGvt39K.

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3D LLM: Learning position-aware video representation for 3d scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8995–9006, 2025.

Yiwu Zhong, Zhuoming Liu, Yin Li, and Liwei Wang. Aim: Adaptive inference of multi-modal llms via token merging and pruning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20180–20192, October 2025.

Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, volume 16, pp. 321–328, 2003. URL https://proceedings.neurips.cc/paper/2003/hash/87682805257e619d49b8e0dfdc14affa-Abstract.html.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13691–13701, 2025. doi: 10.1109/CVPR52734.2025.01278.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. LLaVA-3D: A simple yet effective pathway to empowering LMMs with 3d capabilities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4295–4305, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhu_LLaVA-3D_A_Simple_yet_Effective_Pathway_to_Empowering_LMMs_with_ICCV_2025_paper.html.

A Implementation and Evaluation Details

A.1 Shared Implementation

UG-Pruner prunes visual tokens before the language model while keeping all model parameters frozen. Unless otherwise stated, we retain 25% and 35% of the visual tokens as graph candidates at pruning ratios of 90% and 80%, respectively, and perform five propagation steps. For streaming video, we retain 50% as candidates at a pruning ratio of 75%. Since video inputs contain relatively less redundancy, we additionally adopt a 75% candidate ratio with random candidate selection. The selected tokens keep their original positional indices and are restored to the input order before entering the language model. We set $\tau_s = 0.5$ for image, offline-video, and 3D inputs and $\tau_s = 0.8$ for streaming video, with $\rho_s = 6$, $\Delta_t = 5$, and $r = 1$.

Image. We construct one graph over the image candidates using $C_{\text{sem}} \vee C_{\text{2D}}$ and retain the highest-scoring tokens under the target budget.

Offline Video. We construct a unified graph over candidate tokens from all sampled frames using $C_{\text{sem}} \wedge C_{\text{2D}} \wedge C_{\text{temp}}$. After propagation, global selection allows tokens from different frames to compete for a shared budget and adaptively removes spatio-temporal redundancy across the video.

Streaming video. We sample frames at 0.5 FPS and update the graph every 16 observed frames using the newly arrived candidates and retained history. The constraints $C_{\text{sem}} \wedge C_{\text{2D}} \wedge C_{\text{temp}} \wedge C_{\text{causal}}$ restrict each update to the observed prefix, which is compressed into the fixed visual-token memory.

Multi-view 3D scenes. We use the multi-view inputs and aligned 3D coordinates supplied by the Video-3D-LLM preprocessing pipeline, with $C_{\text{sem}} \vee C_{\text{3D}}$ connecting candidates within or across views. After graph propagation, importance-conditioned farthest point sampling (IFPS) starts from the highest-scoring candidate and iteratively maximizes the product of the propagated score and the nearest 3D distance.

A.2 Cross-Task Baseline Transfer

For cross-task evaluation, image pruning methods process each video frame or 3D view globally and concatenate the retained tokens in their original order. Offline-video methods receive the complete ordered sequence, whereas streaming methods receive frames sequentially and retain their native causal update. Since both video and multi-view 3D tasks use video-form inputs, video pruning methods are directly applied to the ordered 3D views without additional geometric information. All transferred methods preserve their original selection rule and use the same target token budget.

A.3 Evaluation Protocols

We follow the official evaluation protocol of each benchmark. For image understanding, we report accuracy on GQA, MMBench, and TextVQA, F1 on POPE, and the aggregate score on MME. Offline-video evaluation uses accuracy on NExT-QA, LongVideoBench, and MVBench, together with the official aggregate scores of Video-MME and MLVU. For streaming video, we report the real-time, backward, and forward subsets of OVO-Bench, their mean, and the official Streaming-Bench score. Following the 3D evaluation setting used by Geo3DPruner (Li et al., 2026b), we report Acc@0.25/0.5 on ScanRefer, F1@0.25/0.5 on Multi3DRefer, BLEU-4 and CIDEr on Scan2Cap, CIDEr and exact match (EM) on ScanQA, and EM on SQA3D.

B ALGORITHM DETAILS

Algorithm 1 follows the three stages in Figure 2: node selection, constrained graph building, and node estimation.

Algorithm 1 UG-Pruner

Require: Visual token features $\mathbf{F}_v$, candidate budget M, retained-token budget K, task constraints $\mathcal{C}_{\text{task}}$, propagation steps T, optional 3D coordinates $\{\mathbf{z}_i\}$
Ensure: Retained token features $\mathbf{F}'_v$

Node selection
Select M candidate tokens from $\mathbf{F}_v$ as graph nodes $\mathcal{V}$
Constrained graph building
for each node pair $r_{i,j} = (v_i, v_j)$ in $\mathcal{V}$ **do**
    Set $E_{ij} \leftarrow 1$ if $r_{i,j}$ satisfies $\mathcal{C}_{\text{task}}$; otherwise $E_{ij} \leftarrow 0$
    $W_{ij} \leftarrow E_{ij} \dfrac{\mathbf{f}_i^\top \mathbf{f}_j}{\|\mathbf{f}_i\|_2 \|\mathbf{f}_j\|_2}$
end for
$\mathcal{E} \leftarrow \{(v_i, v_j) \mid E_{ij} = 1\}$ and $\mathcal{G} \leftarrow \{\mathcal{V}, \mathcal{E}, W\}$
Node estimation and selection
Row-normalize W as $\widehat{W}$ and initialize $S_i^{(0)} \leftarrow \|\mathbf{f}_i\|_2$
for $t = 0$ to $T - 1$ **do**
    $\mathbf{S}^{(t+1)} \leftarrow \widehat{W}^\top \mathbf{S}^{(t)}$
end for
if $C_{3\text{D}} \in \mathcal{C}_{\text{task}}$ **then**
    Select K nodes by $\text{IFPS}(\mathcal{V}, \mathbf{S}^{(T)}, \{\mathbf{z}_i\}, K)$
else
    Select the top-K nodes by $S_i^{(T)}$
end if
Restore the selected tokens to their original order as $\mathbf{F}'_v$
return $\mathbf{F}'_v$