TOWARDS EFFICIENT AND EFFECTIVE SCENE VISUAL REPRESENTATION FOR 3D SPATIAL SCENE UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite recent advances in multimodal large language models (MLLMs), 3D spatial reasoning remains challenging because spatial evidence is often redundant, occluded, or fragmented across densely sampling views. Existing video-based 3D MLLMs typically adapt video MLLMs to 3D tasks while retaining dense visual tokens, resulting in heavy inference costs. Recent post-hoc token compression methods can alleviate these costs but inevitably compromise performance. In this paper, we propose a unified framework that integrates aggressive visual token compression and multi-teacher distillation into a single stage of 3D task fine-tuning, dubbed **SAVR**. Specifically, we compress visual tokens to a prescribed token budget before they enter the LLM, enabling the model to adapt directly to compact visual tokens. To preserve informative visual cues under aggressive compression, we distill complementary priors from diverse vision foundation models into visual token representations at the final LLM layer. A learnable adaptive weighting mechanism further modulates the contributions of individual teachers to mitigate potentially noisy supervision and facilitate selective knowledge transfer. By coupling token compression with task adaptation and complementary expert distillation, our method leads to compact, task-relevant visual representations without requiring a separate post-hoc compression stage and maintains relatively vanilla performance. Extensive experiments across multiple 3D vision and spatial reasoning benchmarks demonstrate SoTA token efficiency with negligible performance degradation.

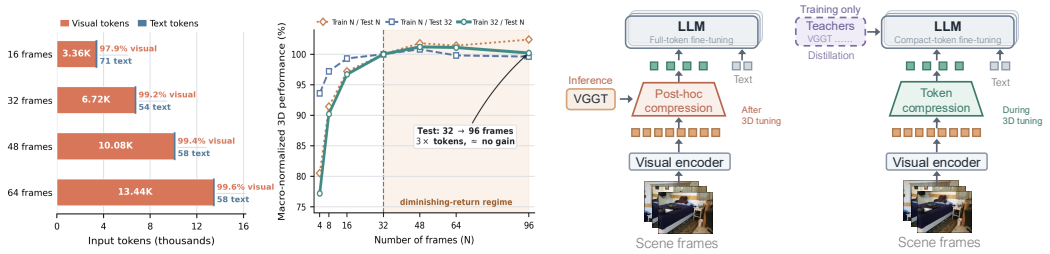
1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have advanced rapidly, demonstrating strong capabilities across a broad range of vision-language tasks (Bai et al., 2025b;a; Hurst et al., 2024; Li et al., 2024; Tong et al., 2024; Zhang et al., 2024c; Team et al., 2024a; Comanici et al., 2025). Building on these advances, recent studies have extended MLLMs to 3D scene understanding domain by encoding point-cloud features (Xu et al., 2024; Chen et al., 2024c;b), lifting multi-view image features into 3D space (Hong et al., 2023; Zhu et al., 2024; Zheng et al., 2025), or incorporating object-centric visual cues (Huang et al., 2024a; Wang et al., 2023; Kang et al., 2025). Despite encouraging progress, representing a 3D scene through multi-views introduces a fundamental efficiency challenge, as reliable spatial reasoning benefits from complete visual coverage across increasing viewpoints.

Existing visual-token reduction methods for general MLLMs estimate token utility from either visual redundancy or cross-modal relevance. Vision-guided approaches merge or discard tokens using feature similarity, `[CLS]` attention, or self-attention (Yang et al., 2025a; Arif et al., 2025; Shang et al., 2025; Wen et al., 2025; Yang et al., 2025c; Zhang et al., 2024a; 2025a; Chen et al., 2024a; Zhang et al., 2025b; Shen et al., 2025; Fu et al., 2024a; Shao et al., 2026), whereas query-guided methods retain tokens most relevant to the instruction (Tan et al., 2025; Zhang et al., 2024b; Xing et al., 2025; 2024; Liu et al., 2024b; Ye et al., 2025; Khaki et al., 2025). Although effective for generic image and video understanding, these criteria primarily capture semantic saliency and redundancy rather than 3D cross- and multi-view correspondences and spatial coverage.

As shown in Fig. 1(a)(b), visual tokens dominate the input sequence, while sampling more views incurs considerable overhead but yields only marginal gains. Recent methods extend token reduction

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) Visual-text token comparison (b) Various sampling frames evaluation (c) Difference from existing works

Figure 1: Preliminary studies of redundancy in 3D MLLM. (a) Number comparisons of visual and text tokens. (b) Pilot experiments of various experiments based on the common 32 input views. (c) Main differences between previous method and our method, without any extra computation overhead of the whole framework during inference and relatively performances retaining.

to 3D MLLMs by exploiting object redundancy, cross-view geometry, or spatial coverage (Huang et al., 2025a; Ling et al., 2026; Huang et al., 2025b; Thomas et al., 2025; Shi & Shin, 2026; Li et al., 2026; Pan et al., 2026; Qu et al., 2026). However, post-hoc methods (Ling et al., 2026; Huang et al., 2025b; Qu et al., 2026) compress visual tokens but induce a train–test mismatch that can degrade performance under aggressive reduction. Geometry-aware compressions (Li et al., 2026; Jiang et al., 2026) require auxiliary heavy geometry models. Although effective, these test-time methods may introduce substantially computational overhead obtained from reducing LLM input sequence, or compromise model performances. These limitations motivate us to explore compact yet spatially informative visual representations during 3D adaptation with no inference overhead.

In this paper, we first propose a unified framework that integrates visual token compression and 3D task adaptation into a single fine-tuning stage, dubbed **SAVR**. Given visual tokens produced by the visual encoder, a simple yet effective adaptive downsampling first reduces them to a compact token set, determined by a prescribed token budget. The LLM is therefore optimized directly on the compact representation instead of dense visual tokens. Aggressive compression, however, may discard spatial details that are essential for 3D scene understanding and spatial reasoning. Inspired by recent works (Huang et al., 2026; Wu et al., 2026; Zheng et al., 2026) that show 3D MLLMs benefit from dense visual priors, we then introduce a multi-teacher distillation strategy during training only to transfer complementary dense visual priors to the model. Specifically, VGGT (Wang et al., 2025) provides geometry-aware and cross-view information, C-RADIO (Heinrich et al., 2025) contributes general dense semantic cues, and Wan2.1 (Wan et al., 2025) supplies spatio-temporal and generative geometry priors. We align these dense teacher representations with the compact visual token set and distill them into the corresponding visual hidden states at the final LLM layer.

Because these teachers are pretrained with different objectives and exhibit heterogeneous feature distributions, equivalently combining their distillation losses can raise conflicts or noisy supervision. We therefore employ a learnable weighting mechanism to associate each teacher with a learnable uncertainty parameter. The mechanism adaptively determines the contribution of informative teachers while suppressing unreliable or incompatible supervision. By coupling token compression with task adaptation and complementary vision foundation model distillation, our method learns compact yet spatially informative visual representations under aggressive token budgets. Extensive experiments across multiple 3D scene understanding and spatial reasoning tasks demonstrate SoTA token efficiency with negligible performance degradation, while substantially boosting inference efficiency.

Our main contributions are summarized as follows:

- We first introduce a unified framework that performs visual token compression jointly with video-to-3D MLLM adaptation, enabling the model to learn directly from compact visual tokens and avoiding a post-hoc compression stage.
- We propose a training-only multi-teacher distillation strategy that transfers complementary geometric, semantic, and spatio-temporal priors from diverse vision foundation models into compact, task-conditioned visual representations.
- We develop a learnable adaptive weighting mechanism to adaptively balance heterogeneous teacher supervision and mitigate noisy or conflicting priors during training only.
- Extensive evaluations demonstrate that our method achieves state-of-the-art token efficiency and maintains relatively vanilla task performance under aggressive token reduction.

2 Related Work

Multimodal Large Language Models (MLLMs). MLLMs (Bai et al., 2025b;a; Hurst et al., 2024; Li et al., 2024; Tong et al., 2024; Zhang et al., 2024c; Team et al., 2024a; Comanici et al., 2025) have achieved significant progress in 2D image and video understanding. However, recent findings (Liu et al., 2024a; Yang et al., 2025b) indicate a critical limitation: current MLLMs still struggle with complex visual spatial reasoning tasks. To address this challenge, some research efforts (Bai et al., 2025b;a; Tong et al., 2024) have focused on enhancing the models' ability to perceive and process spatial relationships through improved model representation. These methods focus on enriching the representation of each individual image, neglecting the 3D geometry information inherent in the continuous frames of a video.

3D Large Multimodal Models. Recently, there has been a growing interest in integrating 3D information into LLMs (Chen et al., 2024c; Hong et al., 2023; Wang et al., 2023; Huang et al., 2023; Xu et al., 2024; Huang et al., 2024a; Zhu et al., 2024; Chen et al., 2024b; Kang et al., 2025), which advances the progress of 3D scene understanding. Existing methods mainly differ in the type of 3D representation they build. A first line of 3D MLLMs grounds language reasoning in point clouds by either directly encoding point-level features or converting 3D detector proposals into object-centric representations (Xu et al., 2024; Chen et al., 2024b;c; Wang et al., 2023; Huang et al., 2023; 2024a). A second line of work lifts multi-view image features into explicit 3D representations. For example, 3D-LLM (Hong et al., 2023) aggregates features from rendered views, while LLaVA-3D (Zhu et al., 2024) constructs 3D patches by pooling 2D features in voxel space; Robin3D (Kang et al., 2025) further improves 3D adaptation through data generation. A complementary line treats multi-view inputs as video sequences (Qi et al., 2026; Zheng et al., 2025; 2026), incorporating positional information in Video-3D LLM (Zheng et al., 2025) or reconstructed BEV maps in GPT4Scene (Qi et al., 2026). Despite their different designs, these methods often rely on explicit point- or voxel-level representations and 3D-specific modules, mainly focusing on enhancing the representation learning by lifting up general MLLMs into 3D tasks, neglecting the efficiency of the model.

Token Compression for Efficient MLLM. Token compression aims to reduce the computational overhead and complexity in transformer visual encoders and MLLMs by merging or pruning redundant visual tokens (Bolya et al., 2022; Chen et al., 2024a; Kong et al., 2025). Image-based methods are typically either text-agnostic, identifying redundancy through visual similarity or internal attention (Yang et al., 2025a; Arif et al., 2025; Shang et al., 2025; Wen et al., 2025; Yang et al., 2025c; Zhang et al., 2024a; 2025a;b), or text-guided, retaining tokens relevant to the instruction (Tan et al., 2025; Zhang et al., 2024b; Xing et al., 2025; 2024; Liu et al., 2024b; Ye et al., 2025; Khaki et al., 2025). Video-based methods additionally reduce temporal redundancy through clip merging, spatio-temporal fusion, progressive pruning, or dynamic KV-cache reduction (Shen et al., 2024; 2025; Tao et al., 2025; Huang et al., 2024b; Fu et al., 2024a). Recent approaches extend token reduction to 3D MLLMs by exploiting object redundancy, cross-view geometry, or spatial coverage (Huang et al., 2025a; Ling et al., 2026; Huang et al., 2025b; Thomas et al., 2025; Shi & Shin, 2026; Li et al., 2026; Pan et al., 2026; Qu et al., 2026). However, these post-hoc methods either require auxiliary geometry models or rely on complex selection procedures at inference, potentially causing performance degradation or offsetting efficiency gains. In contrast, we unify compression with 3D adaptation, inducing no additional inference overhead and substantially retaining vanilla performance.

3 Method

3.1 Pilot Observation

Video-based 3D MLLMs employ 3D scene understanding and spatial reasoning tasks by lifting video MLLMs, sampling multi-view video frames greedily to enable scene coverage, but resulting in an explosion of visual tokens. As shown in Fig. 1(a), 16, 32, and 64 frames generate 3.36K, 6.72K, and 13.44K visual tokens, accounting for 97.9%, 99.2%, and 99.6% of the entire token sequence, respectively. Increasing the number of views therefore seriously increases MLLM inference overhead. We further conduct pilot studies to investigate whether increasing sampling views can help the model improve performance, particularly for the large or complex scenes. Across all three training and evaluation protocols in Fig. 1(b), performance improves obviously with the first few views but gradually saturates at approximately 32 frames. Increasing the input views uniformly from 32 to 96 frames significantly enlarges visual token numbers but with almost no performance gain. The primary bottleneck is possibly not insufficient visual input, but the low information den-

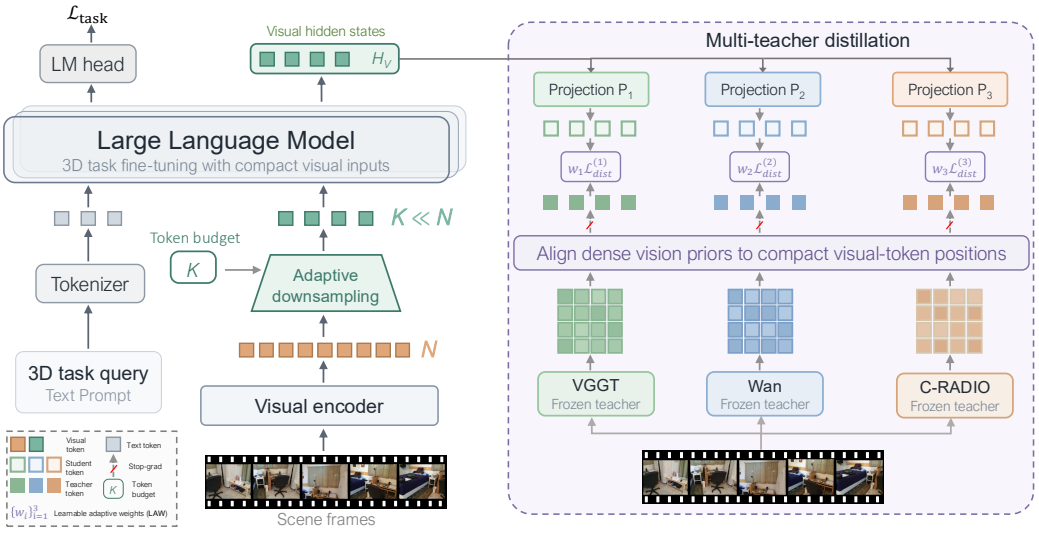


Figure 2: Overall pipeline of our proposed **SAVR**. Our method compresses tokens of video-based 3D MLLMs using a simple yet effective adaptive downsampling strategy, and then distills complementary dense visual priors from vision foundation models as teacher models during training only, which leads to a simple model framework that effectively construct compact visual tokens.

sity of the increased tokens. This observation motivates us to explore *efficiently compact token sets* that perform compression during 3D task adaptation, encouraging compact tokens to consolidate informative evidence and suppress redundant or task-irrelevant content.

3.2 Preliminaries

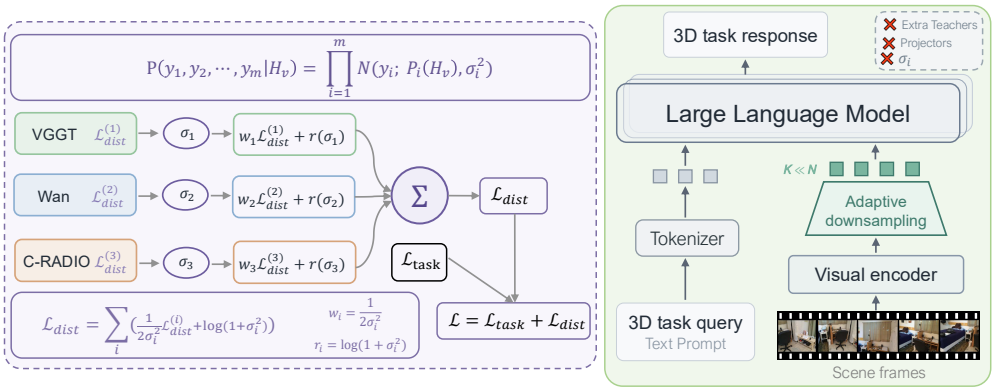
In this paper, we adopt Video-3D LLM (Zheng et al., 2025) as the base model to demonstrate the effectiveness of our token compression framework. Given a sequence of RGB video frames $\{\mathbf{I}_s \in \mathbb{R}^{H\times W\times 3}\}_{s=1}^{M}$, the visual encoder $\mathcal{E}v$, (*i.e.*, SigLIP (Zhai et al., 2023)), tokenizes each frame into patches of size p and produces a corresponding grid of visual embeddings $\mathbf{V} = \{\mathbf{V}_s \in \mathbb{R}^{H_p\times W_p\times d}\}_{s=1}^{M}$, where $H_p = \lfloor H/p \rfloor$, $W_p = \lfloor W/p \rfloor$, and d denotes the feature dimension. A multimodal projector $\mathcal{P}v$ then projects the visual embeddings into LLM embedding space $\mathcal{G}\theta$, (*i.e.*, Qwen2 (Team et al., 2024b)). To construct a 3D-aware video representation, we follow Video-3D LLM to transform depth information into 3D coordinates and integrate these with the visual embeddings using 3D positional encodings for 3D scene understanding. For the spatial reasoning task, we build upon VG LLM (Zheng et al., 2026) without 3D-aware position encoding since the spatial dataset lacks corresponding depth and camera pose annotations.

Although effective, Video-3D LLM incurs a substantial computational overhead, as each frame generates hundreds of visual tokens, resulting in thousands of tokens after sampling video frames. This visual token burden motivates our proposed framework **SAVR**, as shown in Fig. 2. Given a sequence of scene frames, the visual encoder first embeds video frames, which are reduced by an adaptive downsampling module to compact tokens under a prescribed budget. The compact visual tokens are then fed into the LLM for task-specific fine-tuning with a 3D task in Sec. 3.3. The visual hidden states at the final LLM layer will be aligned to distill the complementary multi-teacher priors with a learnable adaptive weighting mechanism in Sec. 3.4. During inference, all teacher models and distillation-specific components are removed, leaving only the vanilla MLLM pipeline to process compact visual tokens efficiently in Sec. 3.5.

3.3 Budget-Adaptive Visual Token Compression

Budget-Constrained Spatial Token Pooling. Given sampling scene frames $\{\mathbf{I}_s \in \mathbb{R}^{H\times W\times 3}\}_{s=1}^{M}$, the visual encoder $\mathcal{E}v$ produces visual feature embeddings $\mathbf{V} = \{\mathbf{V}_s \in \mathbb{R}^{H_p\times W_p\times d}\}_{s=1}^{M}$, where d is the feature dimension and $(H_p \times W_p)$ denotes the spatial token grid of each frame. To constrain the number of visual tokens before LLM processing, we apply parameter-free adaptive average pooling simply to each frame:

$$\widetilde{\mathbf{V}} = \text{AdaptiveAvgPool2d}_{R\times R}(\mathbf{V}) \in \mathbb{R}^{M\times D\times R\times R}. \tag{1}$$



(a) Learnable adaptive weights (LAW) (b)Efficient inference

Figure 3: Illustration of our proposed (a) learnable adaptive weighting formulation paradigm and (b) our final effective and efficient inference framework.

where R is the prescribed output resolution conditioned on the token budget. The original visual sequence contains $N = MH_pW_p$ tokens, whereas the pooled representation now contains $K = MR^2$ tokens, corresponding to a compression ratio:

$$\rho = \frac{K}{N} = \frac{R^2}{H_pW_p}. \tag{2}$$

Specifically, the output feature at spatial position (i, j) of frame s is computed as,

$$\widetilde{\mathbf{v}}_{s,i,j} = \frac{1}{|\mathcal{I}_i^H||\mathcal{I}_j^W|} \sum_{u\in\mathcal{I}^H i} \sum v \in \mathcal{I}^W j \mathbf{v}_{s,u,v}, \tag{3}$$

where $(\mathcal{I}^H i)$ and $(\mathcal{I}^W j)$ denote the adaptively determined pooling intervals along the height and width, respectively. Their boundaries follow the common adaptive average pooling:

$$a^H i = \left\lfloor \frac{(i-1)H_p}{R} \right\rfloor + 1, \quad b^H i = \left\lceil \frac{iH_p}{R} \right\rceil, \; a^W j \quad = \left\lfloor \frac{(j-1)W_p}{R} \right\rfloor + 1, b^W j \quad = \left\lceil \frac{jW_p}{R} \right\rceil. \tag{4}$$

The pooled feature maps are subsequently flattened in frame-first order to form $(\widetilde{\mathbf{V}}_{\text{seq}} \in \mathbb{R}^{K\times d})$ and fed into the LLM. This operation makes no modification to the model architecture and forces the model to abstract essential spatial and object representations from the compact token set. It introduces no complicated token selection or additional trainable parameters.

3.4 Multi-Teacher Prior Distillation

Prior-to-token alignment. Different vision foundation models vary in input resolutions or patch sizes, producing heterogeneous spatial token grids or channel dimensions. Let the dense representation of the i-th teacher be $\mathbf{T}^{(i)} \in \mathbb{R}^{M\times d_i\times H_i\times W_i}$. To align with the compact spatial tokens of the student, we apply the same adaptive downsampling operation to each teacher:

$$\widetilde{\mathbf{T}}^{(i)} = \text{AdaptiveAvgPool2d}_{R\times R}\left(\mathbf{T}^{(i)}\right) \in \mathbb{R}^{M\times d_i\times R\times R}. \tag{5}$$

The pooled teacher features are then rearranged and flattened in frame-first order:

$$\overline{\mathbf{T}}^{(i)} = \text{Flatten}\left(\widetilde{\mathbf{T}}^{(i)}\right) \in \mathbb{R}^{K\times d_i}, \qquad K = MR^2. \tag{6}$$

Consequently, both the teacher and student contain R^2 position-aligned spatial tokens for each frame. Although their token positions are aligned, the channel dimension d_i generally varies across teachers and differs from the student hidden dimension d_s. We therefore introduce a separate lightweight student projector $P_i : \mathbb{R}^{d_s} \to \mathbb{R}^{d_i}$ for each teacher. Given the final-layer student visual hidden states $\mathbf{H}_v \in \mathbb{R}^{K\times d_s}$, the teacher-specific student representation is computed as $\widehat{\mathbf{H}}_v^{(i)} = P_i\left(\mathbf{H}_v\right) \in \mathbb{R}^{K\times d_i}$. We then perform token-wise distillation between the student and teacher as follows:

$$\mathcal{L}_{\text{dist}}^{(i)} = \ell_i\left(\widehat{\mathbf{H}}_v^{(i)}, \text{sg}\left(\overline{\mathbf{T}}^{(i)}\right)\right), \tag{7}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. The teacher models always remain frozen, while the teacher-specific student projectors are jointly optimized with the model. In this way, heterogeneous dense teacher priors are aligned with the compact student token set to minimize information loss.

Learnable Adaptive Weights The heterogeneous yet imperfect teachers provide complementary visual priors but may differ substantially in feature scale, supervision uncertainty, and compatibility with the compact student representation. Therefore, assigning handcrafted fixed weights to all distilled models may induce noisy or conflicting signals that confuse the model training. To this end, we further introduce a learnable adaptive weighting (LAW) mechanism in Fig. 3(a) that associates each teacher with an uncertainty parameter (Kendall et al., 2018). Specifically, let $\mathbf{T}^{(i)}$ denote the aligned target features of the i-th frozen teacher, and let $P_i(\mathbf{H}_v)$ denote the corresponding student representation obtained after projecting the final-layer visual hidden states $\mathbf{H}_v$. We model each teacher target as a Gaussian-likelihood formulation of the projected student representation:

$$p\left(\mathbf{T}^{(1)}, \ldots, \mathbf{Y}^{(p)} \mid \mathbf{H}_v\right) = \prod_{i=1}^{P} \mathcal{N}\left(\mathbf{T}^{(i)}; P_i(\mathbf{H}_v), \sigma_i^2 \mathbf{I}\right). \tag{8}$$

where P denotes the number of teachers and σ_i is a learnable scalar characterizing the uncertainty associated with i-th teacher. In our implementation, $P = 3$ corresponds to VGGT, Wan2.1, and C-RADIO, and uncertainty parameters are jointly optimized with the student model. Motivated by this uncertainty-aware formulation, we define the learnable adaptive weighting objective as:

$$\mathcal{L}_{\text{LAW}} = \sum_{i=1}^{P} \left(\frac{1}{2\sigma_i^2} \mathcal{L}_{\text{dist}}^{(i)} + \log\left(1 + \sigma_i^2\right) \right). \tag{9}$$

where $\mathcal{L}_{\text{dist}}^{(i)}$ is the distillation loss for the i-th teacher. Equivalently, the contribution of each teacher can be expressed as:

$$w_i \mathcal{L}_{\text{dist}}^{(i)} + r(\sigma_i), \qquad where \quad w_i = \frac{1}{2\sigma_i^2}, \qquad r(\sigma_i) = \log\left(1 + \sigma_i^2\right). \tag{10}$$

A smaller learned uncertainty σ_i produces a larger effective weight w_i, strengthening supervision from a teacher that is more conducive to the current student distillation and *vice versa*. The regularization term $r(\sigma_i)$ prevents the model from trivially increasing σ_i to suppress the corresponding distillation objective. Together, these two terms allow the relative contributions of heterogeneous teachers to be learned automatically without manually designing loss coefficients of each teacher.

Finally, the overall training objective consists of the original 3D task loss and the adaptively weighted distillation loss:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{dist}} \mathcal{L}_{\text{LAW}}. \tag{11}$$

where λ_{dist} controls the overall strength of multi-teacher supervision and is set to 1.0 by default. The aligned teacher targets are detached from the computation graph; therefore, gradients from $\mathcal{L}_{\text{LAW}}$ update only the student model, teacher-specific projection layers, and learnable uncertainty parameters. During inference, all frozen teachers, projection layers, and adaptive weighting components are removed, leaving only the compact 3D MLLM for efficient prediction.

3.5 Efficient Inference

During inference, all training-only components, including the frozen vision teachers, teacher-specific projectors, and learnable uncertainty parameters, are completely removed. Given the scene frames, the visual encoder features are directly reduced to compact visual tokens for LLM processing. Since the model is fine-tuned with the same compact representation, our approach avoids the train–test mismatch of post-hoc pruning methods. Moreover, unlike geometry-aware approaches that require auxiliary encoders or costly token selection during inference, our method introduces no extra module overhead while substantially retaining the vanilla performance shown in Fig. 3(b).

4 Experiment

4.1 Datasets and Metrics

ScanNet (Dai et al., 2017) is a large-scale RGB-D video dataset consisting of 1,513 indoor scenes that each scene is annotated with 3D reconstructions, camera poses, depth maps, and object-level semantic segmentation, making ScanNet a widely used benchmark for 3D scene understanding across 3D visual grounding, 3D dense captioning, and 3D question answering tasks.

3D Scene Understanding. For 3D visual grounding, we evaluate our method on ScanRefer (Chen et al., 2020) and Multi3DRefer (Zhang et al., 2023), and report Intersection over Union (IoU) thresholds of 0.25 and 0.5. For 3D dense captioning, we use the Scan2Cap (Chen et al., 2021) benchmark,

Table 1: Performance comparison with previous visual token pruning methods when the video sequence length is set to 32. *Avg.* represents the average percentage of performance maintained at the corresponding reduction ratio across five benchmarks and nine metrics. Performance comparison on Video-3D-LLM under different token retention conditions. Avg. Tokens denotes the average number of vision tokens across LLM layers, and TFLOPS denotes the computational cost during LLM inference. Relative Score refers to the average percentage of performance retained across 5 benchmarks compared to the original LLM. † denotes original paper report.

Method	Avg. Tokens↓	TFLOPS↓	ScanRefer		Multi3DRef		Scan2Cap		ScanQA		SQA3D	Relative Score↑
			Acc@0.25	Acc@0.5	F1@0.25	F1@0.5	B-4@0.5	C@0.5	C	EM	EM	
Upper Bound, 100% Tokens												
Vanilla	100.0%	95.66	58.00	51.40	57.17	51.95	39.63	80.41	103.23	29.75	59.81	100%
Light Compression (High Budget, ∼50% Tokens)												
FastV (ECCV24)	53.6%	49.78	57.29	50.88	56.17	51.14	38.73	75.54	100.06	29.02	58.56	97.61%
VFlowOpt (ICCV25)	51.4%	51.29	57.35	51.09	56.91	51.76	40.35	79.01	100.10	29.58	58.06	99.00%
VisPruner (ICCV25)	50.0%	48.16	57.28	50.94	56.19	50.67	39.90	76.82	98.88	29.22	57.94	97.86%
VisionTrim (ICLR26)	54.9%	59.90	57.23	50.69	55.69	50.61	37.08	68.14	96.79	28.77	58.80	95.47%
HiSC (ACM MM26)	**49.2%**	**47.93**	57.90	51.47	56.82	51.67	40.53	**80.57**	101.21	**29.95**	58.37	99.73%
SAVR	51.0%	49.18	**61.20**	**54.51**	**59.33**	**53.96**	**40.73**	80.10	**101.30**	28.71	**59.70**	**101.78%**
Medium Compression (Mid Budget, ∼25% Tokens)												
FastV (ECCV24)	26.8%	26.80	56.64	50.26	53.47	48.85	36.47	66.22	96.56	28.25	57.06	93.48%
VFlowOpt (ICCV25)	26.1%	29.15	56.75	50.50	55.37	50.43	39.40	73.40	96.30	28.45	57.49	96.20%
VisPruner (ICCV25)	25.0%	24.46	56.70	49.97	54.81	49.05	38.10	70.68	96.63	28.66	57.09	94.97%
VisionTrim (ICLR26)	34.6%	37.71	56.05	49.81	52.07	47.54	35.35	61.05	90.67	26.50	57.34	90.45%
HiSC (ACM MM26)	**24.7%**	**24.06**	57.20	50.89	55.75	50.77	39.56	76.01	**100.78**	**29.82**	57.77	97.96%
SAVR	25.0%	24.49	**60.04**	**53.46**	**58.31**	**53.15**	**40.43**	**79.67**	99.48	28.70	**58.39**	**100.38%**
Extreme Compression (Low Budget, ∼10% Tokens)												
FastV (ECCV24)	10.7%	15.84	55.05	49.12	48.03	44.04	33.31	50.84	85.27	24.51	54.20	84.68%
VFlowOpt (ICCV25)	10.3%	15.34	55.52	49.30	50.64	46.24	35.95	60.19	90.46	26.55	54.96	89.28%
VisPruner (ICCV25)	10.0%	10.18	55.24	48.73	52.06	46.53	36.73	62.44	90.67	26.50	55.84	90.14%
VisionTrim (ICLR26)	19.8%	24.38	54.33	48.55	45.45	41.94	32.44	50.63	83.06	24.04	54.16	82.78%
Geo3DPruner (CVPR26)†	10.2%	10.44	58.00	51.80	53.90	49.30	**41.30**	**82.30**	92.10	26.80	55.80	96.57%
HiSC (ACM MM26)	9.8%	9.9	56.01	49.82	52.40	47.71	37.59	67.95	96.21	28.36	56.27	93.22%
SAVR	**8.2%**	**8.5**	**59.59**	**52.87**	**58.11**	**52.78**	39.80	77.37	**98.97**	**28.47**	**59.59**	**99.63%**

by evaluating traditional image captioning metrics, such as BLEU-4 (Papineni et al., 2002) and CIDEr (Vedantam et al., 2015), with an IoU threshold of 0.5 between prediction and ground-truth bounding boxes. For 3D question answering, we use ScanQA (Azuma et al., 2022) for spatial reasoning and report CIDEr (Vedantam et al., 2015) and exact match (EM) as evaluation metrics, while utilizing SQA3D (Ma et al., 2022) for situated reasoning and adopting exact match (EM).

Spatial Reasoning. Following VG LLM (Zheng et al., 2026), we train our model on a dataset sampled from SPAR-7M (Zhang et al., 2026) and the LLaVA-Hound split of the LLaVA-Video-178K (Zhang et al., 2024c). We first evaluate our method on a video spatial reasoning benchmark, VSI-Bench (Yang et al., 2025b). Additionally, we test our model on an image spatial-related benchmark *i.e.,* CV-Bench (Tong et al., 2024), which assesses 2D understanding through spatial relationships and object counting, while its 3D evaluation focuses on depth ordering and relative distance perception. We also conduct tests on BLINK (Fu et al., 2024b; Ray et al., 2024) to evaluate the spatial subset of the BLINK validation set, which includes three tasks requiring 3D spatial perception abilities, namely relative depth, spatial reasoning, and multi-view reasoning.

4.2 Implementation Details

We adopt Video-3D LLM (Zheng et al., 2025) as our baseline for 3D scene understanding, which is built upon LLaVA-Video 7B (Zhang et al., 2024c). VG LLM (Zheng et al., 2026) is used to construct the baseline for spatial reasoning using Qwen2.5-VL-7B-Instruct (Bai et al., 2025b). All training settings of Video-3D LLM remain consistent with the default configuration in its official open-source codebase, while adding newly trainable parameters, *e.g.,* student projectors and learnable adaptive weights. We utilize VGGT-1B (Wang et al., 2025), Wan2.1-VACE-1.3B (Wan et al., 2025), and C-RADIO_V3 G (Heinrich et al., 2025) as our distillation teachers while remaining frozen to support the vision foundation priors for the 3D MLLM. For each scene, video frames are uniformly sampled from the original sequence. The frames are preprocessed with resizing and center cropping to a resolution of 384×384. During inference, our method is applied to remove redundant visual tokens within the adaptive downsampling mechanism, conditioned on a specified token budget and inducing no extra computational overhead. The loss weighting of overall distillation λ_{dist} is set to 1.0 by default, while L_2 loss is employed for each teacher distillation $\mathcal{L}^i_{dist}$.

Table 2: Evaluation on VSI-Bench with VG LLM as the base model. Under each ratio, the number of tokens retained by each method was kept consistent. Our method consistently exhibits superior performance across various token budgets.

Method	Retention Ratio R	Avg.	Numerical Answer				Multiple-Choice Answer			
			Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
VG LLM + VGGT	-	50.70	67.90	37.70	58.60	62.00	46.60	40.70	32.40	59.20
Vanilla	100%	48.97	68.67	37.02	57.81	62.08	44.78	37.52	31.44	52.42
VisionZip (CVPR25)	25%	43.80	67.39	30.00	50.54	50.32	40.98	33.05	28.89	49.30
FastVID (NeurIPS25)	25%	43.04	64.98	33.21	51.09	49.09	41.20	31.01	30.55	43.19
SAVR	25%	48.92	68.44	34.56	56.18	56.04	45.07	38.86	32.47	59.71
VisionZip (CVPR25)	10%	38.87	64.33	25.00	48.38	46.81	29.02	28.21	24.14	45.08
FastVID (NeurIPS25)	10%	38.52	60.76	26.88	48.97	45.78	30.49	27.35	24.83	43.16
SAVR	10%	48.22	67.13	34.88	55.44	54.09	45.77	43.82	30.41	54.20

Table 3: Ablation study of the effects of different proposed components. All models are fine-tuned with the same training data under 8.2% token retention ratio.

Type	ScanRefer		Multi3DRefer		Scan2Cap		ScanQA		SQA3D
	Acc.25	Acc.5	F1.25	F1.5	C.5	B-4.5	C	EM	EM
Vanilla	58.00	51.40	57.17	51.95	39.63	80.41	103.23	29.75	59.81
Baseline (*w* Adaptive Sampling Only)	57.01	50.60	56.25	50.56	38.10	73.16	99.67	28.06	58.05
+ VGGT	60.04	53.06	58.31	53.10	38.43	73.23	98.48	28.10	58.30
+ VGGT + Wan2.1	59.16	52.24	58.01	51.91	39.30	76.03	99.55	28.91	58.44
+ VGGT + Wan2.1 + C-RADIO	59.04	52.72	57.91	52.31	39.52	77.88	100.01	28.52	58.12
+ VGGT + Wan2.1 + C-RADIO + LAW	59.59	52.87	58.11	52.78	39.80	77.37	98.97	28.47	59.59

4.3 Experimental Results and Analysis

We conduct a comprehensive evaluation of our method by comparing it with existing visual token pruning methods under varying pruning ratios. Note that previous methods usually employ either post-hoc visual token compression strategies or adopt auxiliary geometric encoders, *e.g.,* VGGT, to provide 3D geometry constraints, which then introduces substantial computational overhead. Tab. 1 reports the performance of different pruning strategies when retaining 3,200, 1,568, and 512 tokens with videos consisting of 32 frames, corresponding to pruning ratios of 50%, 75%, and 90%, respectively. At a moderate pruning ratio of 50%, our method has already surpassed the original performance, achieving 101.78% of the baseline accuracy. When the pruning ratio increases to 75%, our method still maintains no vanilla performance drop, with 100.38% relative to the unpruned model. Even under extreme pruning, retaining only 8.2% of the visual tokens, our method achieves 99.63% of the baseline accuracy, substantially outperforming both the text-guided method FastV (Chen et al., 2024a), vision-guided methods VisPruner (Zhang et al., 2025b) and VisionTrim (Yu et al., 2026), and the geometric encoder-guided Geo3DPruner (Li et al., 2026), with negligible performance degradation. These results highlight the effectiveness of our unified token compression distillation design, which *leverages complementary vision foundation models to alleviate the information loss*, thereby preserving the informative and structural integrity of 3D spatial contexts.

4.3.1 Results on 3D Spatial Reasoning.

Tab. 2 presents results on VSI-Bench(Yang et al., 2025b). VG LLM applies the feature fusion with VGGT-based geometric encoder during both training and testing to inject geometric-awareness for visual features while ours distills complementary visual priors and can be discarded after training. Compared with base model, our method consistently sets the highest average scores among token compression methods under the same token budgets, outperforming all baselines including attention-diversity fused strategies and marginal performance drops. We report the results on CV-Bench (Tong et al., 2024) in Tab. 4 and BLINK (Fu et al., 2024b; Ray et al., 2024) in Tab. 5 to evaluate the spatial subset of the BLINK validation set. Under the aggressive compression ratio of 90%, our method can still retain 98.5% and 102.2% of the vanilla model performance, respectively. Interestingly, we find that our method obtains better results than the base model even under 90% token compression while being superior to the base model under a 25% ratio for BLINK spatial-aware evaluation. This

further demonstrates the promising effectiveness and efficiency of our unified framework for scene abstraction visual representation learning.

4.4 Ablation Studies

For this ablation, we performed comprehensive ablation studies to analyze the contributions of each component in the proposed SAVR. All experiments use 32-frame videos with a 90% pruning ratio. We report results on five representative benchmarks: Acc@0.5 and Acc@0.25 on ScanRefer and F1@0.5 and F1@0.25 on Multi3DRefer for 3D visual grounding, BLEU-4 and CIDEr@0.5 on Scan2Cap for 3D dense captioning, and CIDEr on ScanQA and EM on SQA3D for 3D question answering.

Model	2D (%)	3D (%)	Avg. (%)
$VGLLM_{+VGGT}$	72.2	91.1	81.7
Vanilla	74.9	88.3	81.6
Ours *w* 25%	74.7	88.1	81.4
Ours *w* 10%	73.5	87.4	80.4

Table 4: The comparison with state-of-the-art methods on CV-Bench.

Model	MV Rea. (%)	Rel. Depth (%)	Spatial Rea. (%)	Avg. (%)
$VGLLM_{+VGGT}$	-	-	-	51.5
Vanilla	23.3	72.5	76.9	57.6
Ours *w* 25%	**30.1**	68.5	**79.0**	**59.2**
Ours *w* 10%	29.8	68.1	79.0	58.9

Table 5: Comparison of model performance on generic multimodal benchmarks, focusing on spatial parts.

Comparison of adaptive downsampling only. To evaluate the impact of token reduction within adaptive downsampling after the visual encoder, we conduct the preliminary baseline shown in Tab. 3 in the second row. Compared with the vanilla model, applying adaptive downsampling only to compress the visual tokens into a specified budget will drastically degrade the performance since many crucial spatial and semantic details are lost, leading to around 96.3% of the vanilla performance.

Impact of each teacher prior distillation. Since diverse vision foundation models adopt various pre-training objectives, input resolution or patch size which lead to heterogeneous yet complementary supervision signals. For example, VGGT provides geometry-aware and cross-view information, C-RADIO contributes general dense semantic cues, and Wan2.1 supplies spatio-temporal and generative geometry priors. In this section, we conduct experiments by gradually adding each of the selected teacher models to showcase their improvements. As shown in Tab. 3, adding VGGT distillation into the compact token set after adaptive downsampling can significantly improve referring tasks (*e.g.,* ScanRefer and Multi3DRefer). After introducing both Wan2.1 and C-RADIO further boost the semantic and global-scene understanding tasks (*e.g.,* ScanQA and Scan2Cap), improving from roughly 96.3% $\rightarrow$ 98.4% $\rightarrow$ 98.9% $\rightarrow$ 99.2% in terms of vanilla performance.

Improvement of LAW mechanism. Because vision foundation models usually exhibit heterogeneous feature distributions, equivalently combining their distillation losses can raise conflicts or noisy supervision. We further conduct experiments between *w* and *w/o* ablations in Tab. 3. We can see that integrating multi-teacher distillation consistently improves the performance from 99.2% to 99.6%, demonstrating its effectiveness in adaptively balancing the heterogeneous supervisions.

4.5 Efficiency Analysis

In the performance-efficiency studies for 32-frame videos shown in Fig. 1 (a)(b) and Tab. 1, our method consistently outperforms all existing token compression methods for video-based 3D MLLMs, *e.g.,* FastV and VisPruner. Moreover, regarding visual token retention ratio, TFLOPS, and relative scores, ours introduces no additional computation overhead during inference, while Geo3DPruner relies on a VGGT-1B geometric encoder to compute 3D constraints. Our method is capable of obtaining negligible performance degradation even under an aggressive token budget of 8.2%. This observation indicates that our method indeed removes redundant or noisy visual tokens and produces a more compact representation of the 3D scene.

5 Conclusion

In this paper, we first propose to unify visual token compression and 3D adaptation from video MLLMs to construct effective and efficient scene abstraction visual representation learning by using adaptive-weighting multi-teacher distillation. We obtain a compact token set to enforce the model to focus on essential visual cues and learn to abstract the visual priors from teachers. In this way, our method leads to competitive vanilla performance retaining without any additional computational overhead during inference, which hopes to shed new light on the community.

AI USE STATEMENT

In this work, we did not use generative AI tools for any of the tasks with required disclosure. Specifically, we have not used generative AI tools to propose or refine hypotheses, develop conceptual frameworks, design or provide feedback on the research methodology or experiments, implement the proposed method, clean or reformat datasets, interpret results, or assist with translation. Generating synthetic datasets, formulating or proving mathematical claims, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools (e.g., ChatGPT5-SOL and Claude Opus-5) for two tasks with recommended disclosure: (1) writing and editing auxiliary code, such as scripts for plotting figures, reorganzing figures, and aggregating evaluation results, and (2) editing author-written text to improve readability and fluency, including grammar and wording. We have reviewed all AI-assisted work. All AI-assisted scripts were read, executed, and checked for correctness by the authors, and their outputs were cross-checked against the raw experiment logs. For text editing, the AI tools were given only drafts written by the authors, and every suggested change was reviewed to ensure that it did not alter the technical content, claims, or results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. Our study does not involve human subjects, crowdsourcing, or the collection of new data. All experiments are conducted on publicly available 3D scene understanding datasets and benchmarks (*e.g.,* ScanNet and the benchmarks derived from it, SPAR-7M and the LLaVA-Hound split of the LLaVA-Video-178K, and VSI-Bench, CV-Bench and BLINK benchmarks), which we use in accordance with their respective licenses and terms of use. These datasets consist of scans of indoor environments, and we do not attempt to extract or infer any personally identifiable information from them. Our method builds on publicly released pretrained models, including SigLip, LLaVA-Video, Video-3D LLM, VGGT, Wan2.1, Qwen2.5-VL, and C-RADIOv3-G, each used under its original license.

The primary goal of this work is efficiency and effectiveness. By compressing visual tokens, our model retains only about 10% of the visual tokens while maintaining performance comparable to the uncompressed baseline. This reduces the computational and energy cost of training and inference for 3D scene understanding. We do not foresee direct negative societal impacts beyond those common to multimodal LLMs. However, 3D scene understanding models may be deployed in embodied agents or indoor monitoring systems, where they could be misused for surveillance of private spaces. Our model may also inherit biases and hallucination tendencies from the pretrained LLM, the vision teachers, and the training data. The indoor scenes in that data cover a limited range of geographic and cultural settings, and compressed visual representations may discard fine-grained details. We therefore caution against deploying the model in safety-critical settings without further evaluation. And We will release our code and models under licenses compatible with those of the underlying models and datasets.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure that our results are reproducible. Sec. 4 describes the proposed method in detail, including the train-time visual token compression under a given token budget and the distillation objective that aligns the LLM's visual hidden states with downsampled patch tokens from VGGT, Wan2.1, and C-RADIOv3-G. Appendix Sec. A provides the full implementation details, including the model architecture, token budgets, loss weights, optimizer settings, learning rate schedules, batch sizes, number of training steps, and the hardware used for training. Appendix Sec. B describes the datasets and all data processing steps, including frame sampling strategy, and the extraction and downsampling of teacher features, as well as the evaluation protocol and metrics for each benchmark. All datasets and pretrained models used in this work are publicly available and are listed in Section 4. We will release our code and model checkpoints upon acceptance.

REFERENCES

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. In *AAAI*, volume 39, pp. 1773–1781, 2025. 1, 3

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19107–19117. IEEE, 2022. 7

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a. 1, 3

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b. 1, 3, 7

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022. 3

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. 6

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 *ECCV*, pp. 19–35. Springer, 2024a. 1, 3, 8

Sijin Chen, Xin Chen, Chi Zhang, Mingsheng Li, Gang Yu, Hao Fei, Hongyuan Zhu, Jiayuan Fan, and Tao Chen. Ll3da: Visual interactive instruction tuning for omni-3d understanding, reasoning, and planning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26418–26428. IEEE, 2024b. 1, 3

Yilun Chen, Shuai Yang, Haifeng Huang, Tai Wang, Runsen Xu, Ruiyuan Lyu, Dahua Lin, and Jiangmiao Pang. Grounded 3d-llm with referent tokens. *arXiv preprint arXiv:2405.10370*, 2024c. 1, 3

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. 6

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1, 3

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2432–2443. IEEE, 2017. 6

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 visual language models. *arXiv preprint arXiv:2501.01986*, 2024a. 1, 3

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*, 2024b. 7, 8

Greg Heinrich, Mike Ranzinger, Hongxu Yin, Yao Lu, Jan Kautz, Andrew Tao, Bryan Catanzaro, and Pavlo Molchanov. Radiov2. 5: Improved baselines for agglomerative vision foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22487–22497, 2025. 2, 7

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 20482–20494, 2023. 1, 3

Haifeng Huang, Yilun Chen, Zehan Wang, Rongjie Huang, Runsen Xu, Tai Wang, Luping Liu, Xize Cheng, Yang Zhao, Jiangmiao Pang, et al. Chat-scene: Bridging 3d scene and large language models with object identifiers. In *Advances in Neural Information Processing Systems*, volume 37, pp. 113991–114017, 2024a. 1, 3

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19424–19434. IEEE, 2025a. 2, 3

Jiangyong Huang, Silong Yong, Xiaojian Ma, Xiongkun Linghu, Puhao Li, Yan Wang, Qing Li, Song-Chun Zhu, Baoxiong Jia, and Siyuan Huang. An embodied generalist agent in 3d world. *arXiv preprint arXiv:2311.12871*, 2023. 3

Wencan Huang, Daizong Liu, and Wei Hu. Fast3d: Accelerating 3d multi-modal large language models for efficient 3d scene understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10935–10944, 2025b. 2, 3

Xiaohu Huang, Hao Zhou, and Kai Han. Prunevid: Visual token pruning for efficient video large language models. *arXiv preprint arXiv:2412.16117*, 2024b. 3

Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. 3drs: Mllms need 3d-aware representation supervision for scene understanding. *Advances in Neural Information Processing Systems*, 38: 67961–67988, 2026. 2, 16

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1, 3

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. *arXiv preprint arXiv:2605.08064*, 2026. 2

Weitai Kang, Haifeng Huang, Yuzhang Shang, Mubarak Shah, and Yan Yan. Robin3d improving 3d large language model via robust instruction tuning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3905–3915. IEEE, 2025. 1, 3

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7482–7491, 2018. 6

Samir Khaki, Junxian Guo, Jiaming Tang, Shang Yang, Yukang Chen, Konstantinos N Plataniotis, Yao Lu, Song Han, and Zhijian Liu. Sparsevila: Decoupling visual sparsity for efficient vlm inference. In *ICCV*, pp. 23784–23794, 2025. 1, 3

Zhenglun Kong, Yize Li, Fanhu Zeng, Lei Xin, Shvat Messica, Xue Lin, Pu Zhao, Manolis Kellis, Hao Tang, and Marinka Zitnik. Token reduction should go beyond efficiency in generative models–from vision, language to multimodality. *arXiv preprint arXiv:2505.18227*, 2025. 3

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 3

Han Li, Zehao Huang, Jiahui Fu, Naiyan Wang, and Si Liu. Geometry-guided 3d visual token pruning for video-language models. *arXiv preprint arXiv:2604.18260*, 2026. 2, 3, 8

Peng Ling, Yingda Yin, Lingting Zhu, Weikai Chen, Shengju Qian, Zeyu Hu, Xin Wang, and Wenming Yang. Coverprune: Coverage-driven token pruning for 3d vlms via optimal transport. In *European Conference on Computer Vision*, pp. 499–516. Springer, 2026. 2, 3

Benlin Liu, Yuhao Dong, Yiqin Wang, Yongming Rao, Yansong Tang, Wei-Chiu Ma, and Ranjay Krishna. Coarse correspondence elicit 3d spacetime understanding in multimodal language model. *arXiv preprint arXiv:2408.00754*, 3, 2024a. 3

Ting Liu, Liangtao Shi, Richang Hong, Yue Hu, Quanjun Yin, and Linfeng Zhang. Multi-stage vision token dropping: Towards efficient multimodal large language model. *arXiv preprint arXiv:2411.10803*, 2024b. 1, 3

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022. 7

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*, pp. 31066–31077, 2026. 2, 3

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002. 7

Zhangyang Qi, Zhixiong Zhang, Ye Fang, Jiaqi Wang, and Hengshuang Zhao. Gpt4scene: Understand 3d scenes from videos with vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 157277–157302, 2026. 3

Jiuhe Qu, Yingping Liang, and Ying Fu. Hisc: Hierarchical spatial clustering token compression for efficient 3d scene understanding. *arXiv preprint arXiv:2608.04610*, 2026. 2, 3

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, et al. Sat: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024. 7, 8

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *ICCV*, pp. 22857–22867, 2025. 1, 3

Kele Shao, Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Holitom: Holistic token merging for fast video large language models. *Advances in Neural Information Processing Systems*, 38:135547–135570, 2026. 1

Leqi Shen, Tianxiang Hao, Tao He, Sicheng Zhao, Yifeng Zhang, Pengzhang Liu, Yongjun Bao, and Guiguang Ding. Tempme: Video temporal token merging for efficient text-video retrieval. *arXiv preprint arXiv:2409.01156*, 2024. 3

Leqi Shen, Guoqiang Gong, Tao He, Yifeng Zhang, Pengzhang Liu, Sicheng Zhao, and Guiguang Ding. Fastvid: Dynamic density pruning for fast video large language models. *arXiv preprint arXiv:2503.11187*, 2025. 1, 3

Shuyao Shi and Kang G Shin. Feeling the space: Egomotion-aware video representation for efficient and accurate 3d scene understanding. *arXiv preprint arXiv:2603.17980*, 2026. 2, 3

Xudong Tan, Peng Ye, Chongjun Tu, Jianjian Cao, Yaoxin Yang, Lin Zhang, Dongzhan Zhou, and Tao Chen. Tokencarve: Information-preserving visual token compression in multimodal large language models. *arXiv preprint arXiv:2503.10501*, 2025. 1, 3

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. Dycoke: Dynamic compression of tokens for fast video large language models. In *CVPR*, pp. 18992–19001, 2025. 3

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024a. 1, 3

Qwen Team et al. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*, 2024b. 4

Hugues Thomas, Chen Chen, and Jian Zhang. Pts3d-llm: Studying the impact of token structure for 3d scene understanding with large language models. *arXiv preprint arXiv:2506.05689*, 2025. 2, 3

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai C Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems*, volume 37, pp. 87310–87356, 2024. 1, 3, 7, 8

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015. 7

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 7

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025. 2, 7

Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*, 2023. 1, 3

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for important tokens in multimodal language models: Duplication matters more. *arXiv preprint arXiv:2502.11494*, 2025. 1, 3

Xianjin Wu, Dingkang Liang, Tianrui Feng, Kui Xia, Yumeng Zhang, Xiaofan Li, Xiao Tan, and Xiang Bai. Generation models know space: Unleashing implicit 3d priors for scene understanding. In *European Conference on Computer Vision*, pp. 237–258. Springer, 2026. 2, 18

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. 1, 3

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Conical visual concentration for efficient large vision-language models. In *CVPR*, pp. 14593–14603, 2025. 1, 3

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. In *European Conference on Computer Vision*, pp. 131–147. Springer, 2024. 1, 3

Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Chendi Li, Jinghua Yan, Yu Bai, Ponnuswamy Sadayappan, Xia Hu, et al. Topv: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model. In *CVPR*, pp. 19803–19813, 2025a. 1, 3

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10632–10643. IEEE, 2025b. 3, 7, 8

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 *CVPR*, pp. 19792–19802, 2025c. 1, 3

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. Atp-llava: Adaptive token pruning for large vision language models. In *CVPR*, pp. 24972–24982, 2025. 1, 3

Hanxun Yu, Wentong Li, Xuan Qu, Song Wang, Junbo Chen, and Jianke Zhu. Visiontrim: Unified vision token compression for training-free mllm acceleration. *arXiv preprint arXiv:2601.22674*, 2026. 8

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pp. 11975–11986, 2023. 4

Ce Zhang, Kaixin Ma, Tianqing Fang, Wenhao Yu, Hongming Zhang, Zhisong Zhang, Yaqi Xie, Katia Sycara, Haitao Mi, and Dong Yu. Vscan: Rethinking visual token reduction for efficient large vision-language models. *arXiv preprint arXiv:2505.22654*, 2025a. 1, 3

Jiahui Zhang, Yurui Chen, Yueming Xu, Ze Huang, Jilin Mei, Chunhui Chen, Yanpeng Zhou, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, et al. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *Advances in Neural Information Processing Systems*, 38, 2026. 7

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. 1, 3

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20857–20867. IEEE, 2025b. 1, 3, 8

Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15179–15179. IEEE, 2023. 6

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. 1, 3

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024c. 1, 3, 7

Duo Zheng, Shijia Huang, and Liwei Wang. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8995–9006. IEEE, 2025. 1, 3, 4, 7

Duo Zheng, Yanyang Li, Liwei Wang, et al. Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. *Advances in neural information processing systems*, 38:20560–20586, 2026. 2, 3, 4, 7, 18

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness. *arXiv preprint arXiv:2409.18125*, 2024. 1, 3

APPENDIX

This appendix is organized as follows. Appendix A provides the full implementation details, including token accounting, teacher feature extraction, and pseudocode of the training procedure. Appendix B describes the datasets, data processing steps, evaluation protocols, and baseline settings. Appendix C discusses limitations.

A IMPLEMENTATION DETAILS

A.1 TRAINING CONFIGURATION

Tab. 6 summarizes the training configurations of both settings. For 3D scene understanding, all settings not listed below follow the default configuration of the official Video-3D LLM codebase. The only newly introduced trainable parameters are the teacher-specific student projectors $\{P_i\}$ and the uncertainty parameters $\{\sigma_i\}$.

Table 6: Training configurations of the two model framework settings.

	3D scene understanding	Spatial reasoning
Base model	Video-3D LLM	VG LLM (w/o VGGT fusion)
LLM backbone	LLaVA-Video-7B	Qwen2.5-VL-7B-Instruct
Visual encoder	SigLIP	Qwen2.5-VL ViT
Number of frames M	32	LLaVA-Hound=[4, 8], SPAR-7M=[1, 5]
Input resolution	384×384	504×392
Pooled resolution R	10 / 7 / 4	9×7 / 5×5
Teachers (frozen)	VGGT-1B, Wan2.1-VACE-1.3B, C-RADIOv3-G	
Distillation loss ℓ_i	L2	
λ_{dist}	1.0	
Initialization of σ_i	1.0	
Trainable modules	Full fine-tuning	
Optimizer	Adam	AdamW
Learning rate / schedule	1e-5 / cosine	
Warm-up ratio	0.03	
Global batch size	16	64
Training epochs / steps	1 epoch	
Precision	bf16	
Hardware	NVIDIA H20 8xGPUs	

A.2 MORE DEMONSTRATION OF OUR MOTIVATION

To further demonstration our motivation of proposd **SAVR**, we provide illustrations in Fig. 4 and Fig. 5. The primary bottleneck is possibly not insufficient visual input, but the low information density of the increased tokens. This observation motivates us to explore *efficiently compact token sets* that perform compression during 3D task adaptation, encouraging compact tokens to consolidate informative evidence and suppress redundant or task-irrelevant content. Note that, recent 3DRS (Huang et al., 2026) also explores VGGT vision foundation model for 3D MLLMs but with full token set. Differenctly, our method aims at obtaining efficient and effective scene visual representation learning within a unify 3D task adaptation and introduces no any extra computational overhead during inference.

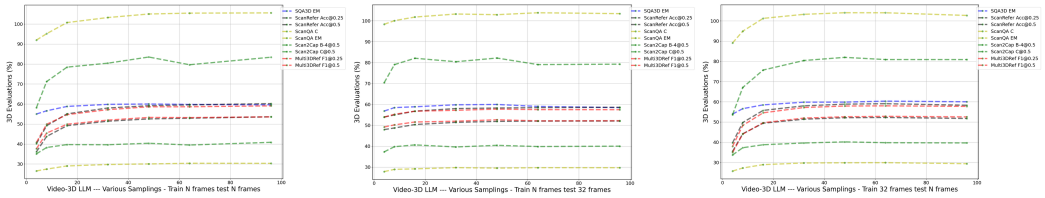
A.3 TOKEN ACCOUNTING AND COMPRESSION RATIO

In the 3D scene understanding setting, each 384×384 frame is encoded by SigLIP and, following LLaVA-Video, pooled to an $H_p \times W_p = 14 \times 14$ grid, *i.e.*, 196 visual tokens per frame. We therefore define $N = MH_pW_p$ with $H_p = W_p = 14$, and the retention ratio $\rho = R^2/(H_pW_p)$ as in Eq. (2) of the main paper. Tab. 1 lists the resulting budgets for $M = 32$ frames, where $N = 6{,}272$.



(a) Various scene reconstruction (b) Neighboring views overlap

Figure 4: VGGT utilization to reconstruct the 3D scene among various sampling views. (a) As can be seen very sparse sampling views cannot enable common objects visible while dense sampling views introduce marginal scene- and object-level understanding capacity but seriously burdening the computational overhead. (b) There are usually multiply overlaps correspondences across neighboring views which raises visual redundancy among visual tokens (higher heatmap visualization denotes more times overlapping projection among pixels across neighboring views). Hence, this motivates our proposed **SAVR** to unify token compression to improve efficiency and visual priors from multi-teachers to maintain negligible performance drop.



Pilot study for various sampling frames evaluation

Figure 5: Detailed pilot studies on various sampling frames training and evaluation, that motivates us to perform compression during 3D task adaptation, encouraging compact tokens to consolidate informative evidence and suppress redundant or task-irrelevant content rather than denser visual tokens covergae.

Table 7: Token budgets for Video-3D LLM with $M = 32$ frames.

Budget	R	$K = MR^2$	Retention ratio ρ
Light	10	3,200	51.0%
Medium	7	1,568	25.0%
Extreme	4	512	8.2%

Per-frame visual token grid. Following VG LLM (Zheng et al., 2026), every input frame is first resized with the VGGT preprocessing: the width is set to 518 pixels, the height is rescaled to preserve the aspect ratio (rounded to a multiple of 14), and the frame is center-cropped to at most 518 pixels in height. For the 2D branch, the frame is further cropped to multiples of 28 pixels, since the Qwen2.5-VL vision encoder uses 14×14 patches and merges every 2×2 neighboring patches into a single visual token, so that each token fed to the LLM covers a 28×28 pixel region. For the 4:3 frames of ScanNet-style scenes, this yields a 504×392 input and an 18×14 token grid, *i.e.*,, $N_f = 252$ visual tokens per frame (16:9 video frames yield a 504×280 input, *i.e.*,, $18 \times 10 = 180$ tokens).

Temporal patch merging. For video inputs, Qwen2.5-VL natively groups every two consecutive frames into a single temporal patch (temporal patch size of 2), which halves the number of visual tokens. We do not apply this temporal merging. Following VG LLM, the sampled frames are fed as a sequence of individual images, each of which is replicated along the temporal axis to fill one temporal patch ($t = 1$ in the (t, h, w) grid). Consequently, every frame keeps its own 18×14 grid, the geometry features can be fused frame by frame, and a clip of T frames produces $252\,T$ visual tokens. During training, LLaVA-Hound clips contain 4–8 frames and SPAR-7M samples contain one or 3–5 images; at evaluation, 32 frames are uniformly sampled for VSI-Bench, resulting in $8{,}064$ visual tokens without compression.

Choice of R. Let R denote the number of visual tokens retained per frame before entering the LLM. We set R as a fixed fraction of N_f so that the retention ratios match the 25% and 10% budgets in Table 2 of the main paper. For the 25% budget, we use $R = N_f/4 = 63$, which corresponds exactly to an additional 2×2 merge of the 18×14 grid (9×7). For the 10% budget, $N_f/10 = 25.2$ is not an integer, and we use $R = 25$ (9.9%). For 16:9 frames, R is scaled proportionally to 45 and 18, respectively. Since R is fixed per frame, the retention ratio is independent of the number of input frames. Tab. 8 summarizes the resulting token counts.

Budget	Tokens / frame (R)	Ratio	32-frame input
Full	252 (18×14)	100%	8,064
25%	63 (9×7)	25.0%	2,016
10%	25	9.9%	800

Table 8: Per-frame and per-sample visual token counts in the spatial reasoning setting (4:3 frames).

A.4 Placement of the Pooling Operation

The adaptive average pooling is applied after the multimodal projector $\mathcal{P}_v$ and before the visual tokens are concatenated with the text tokens. For Video-3D LLM, the 3D positional encodings are computed from the pooled 3D coordinates / pooled together with the visual embeddings, so that each compact token carries the 3D position of the region it summarizes. For VG LLM, the spatial training set has no 3D prior annotations available.

A.5 Teacher Feature Extraction

All teachers are frozen and used only during training. Teacher features are pooled to the same $R \times R$ grid as the student (Eq. (5) in the main paper) and normalized before computing the corresponding $\mathcal{L}^i_{dist}$ loss. Teacher features can be computed online or pre-extracted and cached offline. In this paper, we pre-extract the features to speedup the training. For extracting Wan2.1-VACE-1.3B features, we mainly follow VEGA-3D (Wu et al., 2026). The Flow-Matching time interval $t \in [0, 1]$ is discretized into K=1000 steps in the pretrained Wan2.1 backbone. Then discrete timestep index is set to be k and use $t_k = \frac{k}{K}$ as the normalized time, and extracting features at k=300 (i.e., t_k=0.3) from the 20th DiT layer.

A.6 Student Projectors

Each student projector $P_i : \mathbb{R}^{d_s} \to \mathbb{R}^{d_i}$ is a one-layer MLP with GELU.

A.7 Training Procedure

Algorithm 1 summarizes one training step of the proposed SAVR.

Algorithm 1 One training step of SAVR

Require: Frames $\{\mathbf{I}_s\}_{s=1}^{M}$, query q, target answer y, pooled resolution R, frozen teachers $\{\mathcal{T}_i\}_{i=1}^{P}$, student projectors $\{P_i\}_{i=1}^{P}$, uncertainties $\{\sigma_i\}_{i=1}^{P}$, weight λ_{dist}

1: $\mathbf{V} \leftarrow \mathcal{E}_v(\{\mathbf{I}_s\})$ ▷ visual encoder
2: $\widetilde{\mathbf{V}}_{\text{seq}} \leftarrow \text{Flatten}\big(\text{AdaptiveAvgPool2d}_{R\times R}(\mathbf{V})\big)$ ▷ $K = MR^2$ tokens
3: $\mathbf{H} \leftarrow \mathcal{G}_\theta\big(\mathcal{P}_v(\widetilde{\mathbf{V}}_{\text{seq}}),\, q\big)$ ▷ 3D positional encodings if applicable
4: $\mathcal{L}_{\text{task}} \leftarrow -\log p_\theta(y \mid \widetilde{\mathbf{V}}_{\text{seq}}, q)$
5: $\mathbf{H}_v \leftarrow$ final-layer hidden states at the K visual positions of $\mathbf{H}$
6: **for** $i = 1, \ldots, P$ **do**
7: $\quad \overline{\mathbf{T}}^{(i)} \leftarrow \text{sg}\Big(\text{Flatten}\big(\text{AdaptiveAvgPool2d}_{R\times R}(\mathcal{T}_i(\{\mathbf{I}_s\}))\big)\Big)$
8: $\quad \mathcal{L}_{\text{dist}}^{(i)} \leftarrow \ell_i\big(P_i(\mathbf{H}_v),\, \overline{\mathbf{T}}^{(i)}\big)$
9: **end for**
10: $\mathcal{L}_{\text{LAW}} \leftarrow \sum_{i=1}^{P} \Big(\frac{1}{2\sigma_i^2}\mathcal{L}_{\text{dist}}^{(i)} + \log(1+\sigma_i^2)\Big)$
11: $\mathcal{L} \leftarrow \mathcal{L}_{\text{task}} + \lambda_{\text{dist}}\,\mathcal{L}_{\text{LAW}}$
12: Update θ, $\{P_i\}$, and $\{\sigma_i\}$ by back-propagating $\mathcal{L}$; the teachers stay frozen

At inference time, lines 6–11 are skipped, and the teachers, projectors, and uncertainty parameters can be completely discarded.

B Datasets, Data Processing, and Evaluation

B.1 Training Data

3D scene understanding. We follow the training data of Video-3D LLM, built on the 1,513 ScanNet scans with annotations from ScanRefer, Multi3DRefer, Scan2Cap, ScanQA, and SQA3D; report the number of training samples per task. For each scene, $M = 32$ frames are uniformly sampled from the RGB-D sequence, then resized and center-cropped to 384×384.

Spatial reasoning. Following VG LLM, we train on a subset of SPAR-7M and the LLaVA-Hound split of LLaVA-Video-178K.

B.2 Evaluation Benchmarks and Metrics

Tab. 9 lists the evaluation benchmarks, splits, and metrics.

Table 9: Evaluation benchmarks, splits, and metrics.

Benchmark	Task	Split	Metrics
ScanRefer	3D visual grounding	val	Acc@0.25, Acc@0.5
Multi3DRefer	3D visual grounding	val	F1@0.25, F1@0.5
Scan2Cap	3D dense captioning	val	BLEU-4@0.5, CIDEr@0.5
ScanQA	3D question answering	val	CIDEr, EM
SQA3D	Situated question answering	test	EM
VSI-Bench	Video spatial reasoning	benchmark	MRA / accuracy
CV-Bench	2D and 3D spatial perception	benchmark	Accuracy
BLINK	Spatial subset	benchmark	Accuracy

B.3 Relative Score

For the 3D scene understanding benchmarks, the Relative Score in Tab. 1 of the main paper is the mean ratio over the set $\mathcal{S}$ of the nine reported metrics, namely two metrics each for ScanRefer, Multi3DRefer, Scan2Cap, and ScanQA, and one for SQA3D:

$$\mathrm{RS} = \frac{1}{|\mathcal{S}|} \sum_{m \in \mathcal{S}} \frac{m_{\text{method}}}{m_{\text{vanilla}}} \times 100\%. \tag{12}$$

All methods, including those whose results are taken from the original papers, are normalized by the same vanilla model in Tab. 1.

B.4 Baseline Settings

The training-free baselines (FastV, VFlowOpt, VisPruner, VisionTrim, and HiSC) are applied to the same vanilla Video-3D LLM checkpoint with their official implementations and default hyperparameters, and their budgets are matched to ours as closely as their token selection allows. Results of Geo3DPruner are taken from its original paper (marked with †) since it is not publicly available yet. On VSI-Bench, VisionZip and FastVID are applied to the same vanilla VG LLM under identical retention ratios which is trained with VGGT distillation insted of the oiginal VGGT fusion.

C Limitations

First, our compression is a content-agnostic, per-frame average pooling. It preserves scene coverage across views but cannot allocate more tokens to informative regions, so small objects and fine-grained attributes may be lost under aggressive token budgets. This is consistent with the larger drops on generation-oriented metrics at the 8.2% budget, *e.g.*, CIDEr on Scan2Cap ($80.41 \rightarrow 77.37$) and ScanQA ($103.23 \rightarrow 98.97$), whereas grounding metrics are preserved or improved. Second, although the teachers add no inference cost, running them during training adds computation on top of fine-tuning with compact tokens alone. Third, a separate model is trained for each token budget, so unlike training-free methods, changing the budget requires re-training. Finally, our model may inherit biases from the pretrained LLM, the teachers, and the training data, whose indoor scenes cover a limited range of environments. A promising future work will be to scale up our method into larger and more powerful model options with large-scale training data, so pushing efficient and scalable spatial intelligent model.