TAMP: Token-Adaptive Layerwise Pruning in Multimodal Large Language Models

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) have shown remarkable versatility in understanding diverse multimodal data and tasks. However, these capabilities come with an increased model scale. While post-training pruning reduces model size in unimodal models, its application to MLLMs often yields limited success. Our analysis discovers that conventional methods fail to account for the unique token attributes across layers and modalities inherent to MLLMs. Inspired by this observation, we propose TAMP, a simple yet effective pruning framework tailored for MLLMs, featuring two key components: *(1) Diversity-Aware Sparsity*, which adjusts sparsity ratio per layer based on diversities among multimodal output tokens, preserving more parameters in high-diversity layers; and *(2) Adaptive Multimodal Input Activation*, which identifies representative multimodal input tokens using attention scores to guide unstructured weight pruning. We validate our method on two state-of-the-art MLLMs: LLaVA-NeXT, designed for vision-language tasks, and VideoLLaMA2, capable of processing audio, visual, and language modalities. Empirical experiments across various multimodal evaluation benchmarks demonstrate that each component of our approach substantially outperforms existing pruning techniques.[1]

1 Introduction

Large Language Models (LLMs) have achieved remarkable success at billion-parameter scales (Touvron et al., 2023a,b; DeepSeek-AI et al., 2025), excelling in challenging tasks. Building on this, Multimodal Large Language Models (MLLMs) (Li et al., 2024a; Zhan et al., 2024; Wu et al., 2024), which extend LLMs to handle diverse modality inputs, have grown in size to address the complexities of multimodal tasks (Liang et al., 2024; Tong et al., 2024; Shi et al., 2024). While beneficial

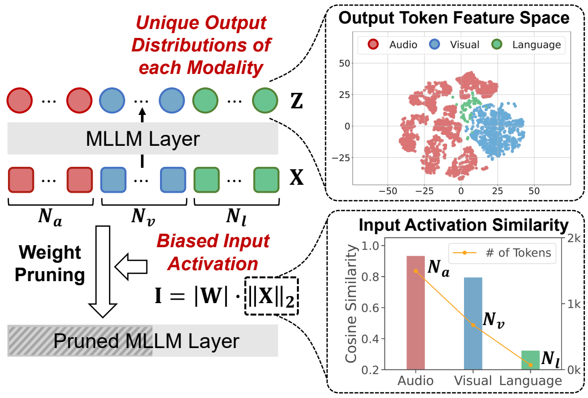


Figure 1: Illustration of multimodal token attributes. **(Top)**: t-SNE visualization of multimodal output tokens of the layer, exhibiting unique distributions of each modality. **(Bottom)**: Cosine similarity between the ℓ_2-norm of tokens from each modality and all tokens, demonstrating a bias in input activations toward the modality with the largest token count (N_a,$N_v >> N_l$), resulting in suboptimal weight pruning.

for performance, their colossal model size imposes substantial computational and memory resources, limiting their practicality in resource-constrained scenarios (Reid et al., 2024; Li et al., 2024c).

Post-training model pruning (Sun et al., 2024b; Frantar and Alistarh, 2023; Ma et al., 2023; Yu and Xiang, 2023) effectively reduces model size by removing a massive number of parameters without compromising performance. Studies on applying LoRA (Zhang et al., 2024a; He et al., 2024) or quantization (Guo et al., 2024) on top of pruned models have been conducted to further enhance the performance and efficiency of pruning strategies. Although effective, most existing techniques assume unimodal models, limiting their effectiveness in multimodal settings. For example, in Figure 1, our empirical examination shows that conventional unimodal pruning methods, such as Wanda (Sun et al., 2024b) and similar pruning approaches (Zhang et al., 2024c; He et al., 2024; Sung et al., 2024; Yin et al., 2024), fail to generalize to multimodal settings where there can be substantial variances in input token activation and output token distributions across modalities (Liang et al., 2024)

[1] Our code will be publicly available upon publication.

Drawing inspiration from these observations, we introduce **T**oken-**A**daptive **M**ultimodal **P**runing (TAMP), a novel MLLM pruning framework that leverages inherent multimodal token attributes. TAMP comprises two key components: first, we employ a layer-wise sparsity ratio strategy that dynamically adjusts the sparsity ratio per layer, guided by the varying output token distributions. Specifically, we assign lower sparsity ratios to layers exhibiting greater output token variations, ensuring that these layers retain sufficient parameters to encode rich multimodal representations. Second, instead of using all input tokens to compute input activations, we utilize attention scores to identify key multimodal input tokens that account for each layer's unique multimodal processing demands.

We validate our approach in various pruning scenarios using two distinct MLLMs, LLaVA-NeXT (Li et al., 2024a) and VideoLLaMA2 (Cheng et al., 2024), evaluated on diverse multimodal benchmarks. Our layer-wise sparsity ratio strategy, based on varying output distributions, alone outperforms recent layer-wise sparsity approaches like ECoFLAP (Sung et al., 2024) and OWL (Yin et al., 2024), with 4.0% higher performance at 50% sparsity. Moreover, our approach of selecting multimodal tokens for input activations achieves up to 4.1% performance gains over the state-of-the-art LLM pruning method Wanda (Sun et al., 2024b) at 50% sparsity. Combining both strategies further enhances performance, consistently surpassing strong pruning baselines. Our approach shows robustness at high sparsity, where our approach outperforms the second-best baseline with 8.2% higher performance at 70% sparsity. Notably, our approach exclusively uses multimodal token attributes, avoiding the need for resource-intensive gradient or Hessian computations (Frantar and Alistarh, 2023; Sung et al., 2024), supporting its efficiency by leveraging multimodal attributes for effective MLLM pruning.

In summary, our contributions are as follows:

- We conduct comprehensive analyses and ablation studies to identify the importance of multimodal tokens in MLLM pruning. These include extensive analyses of multimodal token distributions across layers and in-depth investigations into their impact on pruning.
- We introduce TAMP, an effective MLLM pruning pipeline that leverages multimodal token attributes to measure layer importance for layer-wise sparsity and computes adaptive input activations for capturing multimodal processing demands at each layer.
- We validate our method on MLLMs that reflect their latest trends, demonstrating its effectiveness in preserving diverse multimodal abilities. Ours consistently outperforms pruning baselines, even at extreme pruning ratios.

2 Related Work

Multimodal Large Language Models Recent advancements in Large Language Models (LLMs), such as LLaMA (Touvron et al., 2023a,b), Qwen (Yang et al., 2024), and DeepSeek (DeepSeek-AI et al., 2025) have achieved remarkable progress in various natural language processing tasks by scaling to billions of parameters. Building on this success, Multimodal Large Language Models (MLLMs) have emerged as a new standard of multimodal models, integrating multiple modalities, including text, image, audio and video, into a unified framework to address complex multimodal challenges in the real world (Li et al., 2024c; Zhan et al., 2024; Wu et al., 2024).

LLaVA-NeXT (Li et al., 2024a) integrates a visual encoder into LLMs and facilitates the understanding of high-resolution images, improving tasks such as visual question answering (Masry et al., 2022; Kembhavi et al., 2016) and visual reasoning (Yue et al., 2024; Liu et al., 2024). Expanding beyond images, MLLMs such as VideoLLaMA2 (Cheng et al., 2024) and LLaVA-OneVision (Zhan et al., 2024) have broadened their potential applications by incorporating other modalities such as audio, video, and interleaved images. However, as MLLMs continue to grow in size, their deployment in resource-constrained environments becomes increasingly challenging.

Model compression To tackle the challenges posed by increasing model scale, model compression techniques have emerged as a critical research area, aiming to optimize model size while maintaining performance (Yao et al., 2022; Wang et al., 2024a; Frantar et al., 2023; Lin et al., 2024). Among these techniques, model pruning has gained prominence by removing redundant parameters or structures that minimally contribute to overall performance (Sun et al., 2024b; Zhou et al., 2021; Ma et al., 2023). Approaches like Wanda (Sun et al., 2024b) utilize weight magnitudes and input activations to compress LLMs, while SparseGPT (Fran-

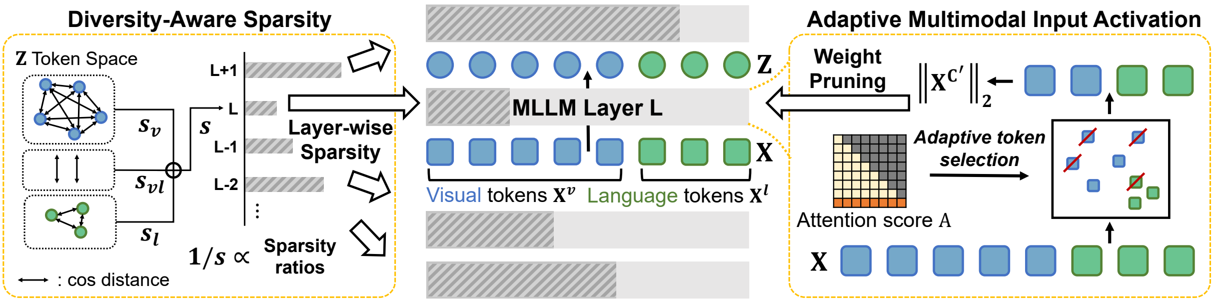


Figure 2: Overview of TAMP. Our method utilizes multimodal token attributes to guide MLLM pruning. (**Left**): To effectively preserve each MLLM layer's differing capability to encode rich multimodal output tokens after pruning, we apply layer-wise sparsity, assigning sparsity inversely to the layer's importance, which is computed as the average of intra-modality ($\mathbf{s}_v, \mathbf{s}_l$) and inter-modality ($\mathbf{s}_{vl}$) diversities (Section 3.2). (**Right**): To capture unique multimodal processing demands across different layers, we leverage attention scores to adaptively select multimodal input tokens for input activation calculations (Section 3.4).

tar and Alistarh, 2023) addresses the challenge of LLM pruning from the perspective of layer-wise output reconstruction problem.

However, all the aforementioned works are primarily designed for unimodal models, limiting their applicability to MLLMs. While ECoFLaP (Sung et al., 2024) and VLMPrune (He et al., 2024) extend pruning strategies to Vision-Language Models (VLMs) by applying layer-wise sparsity ratios tailored to vision-language characteristics, they treat multimodal tokens as if they originate from a single modality, overlooking their unique properties. In contrast, our work examines the impact of multimodal tokens on MLLM weight pruning and explicitly leverages multimodal properties for optimal pruning. Unlike prior approaches, we conduct comprehensive experiments on recent MLLMs, including those with more than two modalities, aligning with the latest advancements in MLLM design.

3 Method

In this section, we first present empirical studies that reveal key properties of multimodal tokens and their implications for pruning. Based on these insights, we introduce the core components of Token Adaptive Multimodal Pruning (TAMP). The overall framework is illustrated in Figure 2.

3.1 Preliminaries

A predominant MLLM typically consists of modality-specific encoders connected to an LLM through intermediate networks, with multimodal information from these encoders provided to the LLM as input tokens. While the following descriptions focus on an MLLM that uses an image encoder for visual information for simplicity, our approach is extensible to MLLMs that process other modalities, such as audio, video, or both.

Each block of the LLM processes two types of input tokens: visual $\mathbf{X}^v \in \mathbb{R}^{N_v \times C_{in}}$ and language $\mathbf{X}^l \in \mathbb{R}^{N_l \times C_{in}}$ input tokens, where N_v and N_l denote their respective token counts, and C_{in} is the input dimension size. The block contains a multi-head attention (MHA) module, which computes an attention score $\mathbf{A} \in \mathbb{R}^{(N_v+N_L)\times(N_v+N_L)}$ that measures interplay between tokens, and a feed-forward network (FFN) module, which refines the output from the MHA module. Within these modules, varying types of linear projection layers $\mathbf{W} \in \mathbb{R}^{C_{out} \times C_{in}}$ transform input tokens into output tokens, where C_{out} is the output dimension size:

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}^v \\ \mathbf{Z}^l \end{bmatrix} = \begin{bmatrix} \mathbf{X}^v \\ \mathbf{X}^l \end{bmatrix} \mathbf{W}^\top = \mathbf{X}\mathbf{W}^\top, \qquad (1)$$

where $\mathbf{Z}^v$ and $\mathbf{Z}^l$ represent the visual and language output tokens, respectively. To determine which parameters to prune, many predominant methods (Sun et al., 2024b; Sung et al., 2024; Yin et al., 2024) define layer's parameter importance based on input activation and weight magnitude, computed as: $\mathbf{I} = ||\mathbf{X}||_2 \cdot |\mathbf{W}|$, where $|\cdot|$ is element-wise absolute value operator and $||\mathbf{X}||_2 \in \mathbb{R}^{C_{in}}$ is the input activation computed for each channel as ℓ_2-norm of that channel's activation across all input tokens. Parameters with the lowest importance are considered redundant and thus pruned.

3.2 Diversity-Aware Sparsity

We first conduct a systemic study of the distributional properties of multimodal output tokens by computing intra- and inter-modality diversities. Intra-modality diversities measure the distances among output tokens within the same modality, while inter-modality diversity quantifies those be-

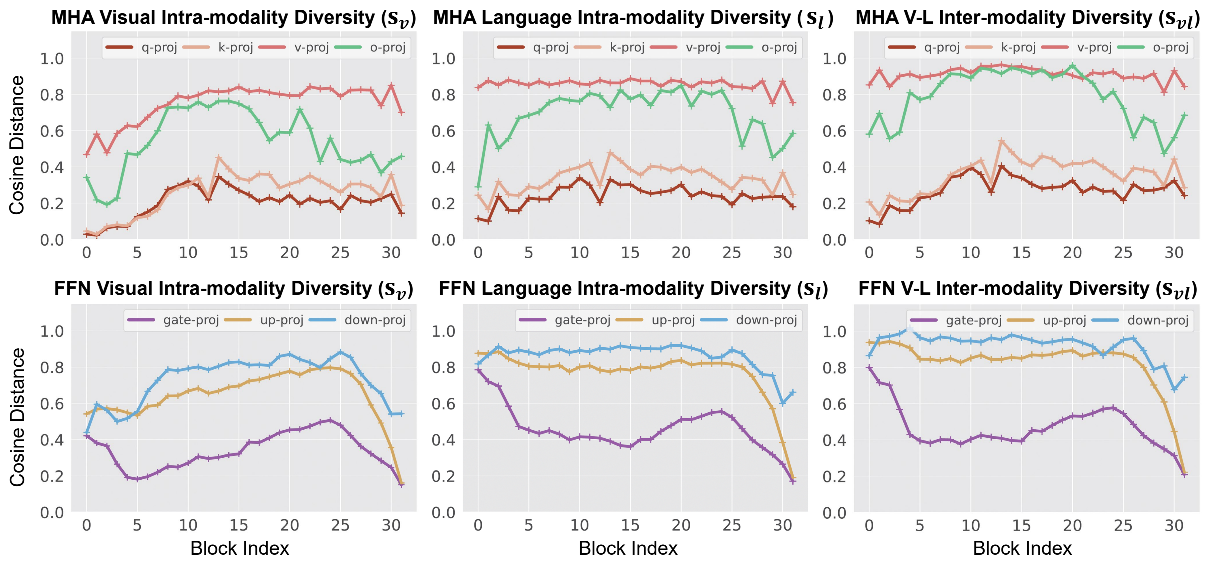


Figure 3: Intra-modality diversities ($\mathbf{s}_v, \mathbf{s}_l$) measure the average cosine distances among output tokens within the same modality, and inter-modality diversity ($\mathbf{s}_{vl}$) measures distances between output tokens from different modalities. We compute these diversities for each projection type in multi-head attention (Top) and feed-forward network (Bottom) across LLaVA-NeXT blocks. Notably, diversity trends differ by (1) modalities, (2) projection types, and (3) blocks, demonstrating varying capacities that should be preserved to effectively encode multimodal information across layers.

tween output tokens from different modalities:

$$\begin{aligned} \mathbf{s}_v &= \mathbb{E}_{i,j\sim\mathcal{C}_v}\left[\mathbf{d}_{ij}\right], \mathbf{s}_l = \mathbb{E}_{i,j\sim\mathcal{C}_l}\left[\mathbf{d}_{ij}\right], \\ \mathbf{s}_{vl} &= \mathbb{E}_{i\sim\mathcal{C}_v,\, j\sim\mathcal{C}_l}\left[\mathbf{d}_{ij}\right], \mathbf{d}_{ij} = 1 - \langle \mathbf{Z}_i, \mathbf{Z}_j \rangle, \end{aligned} \tag{2}$$

where $\mathcal{C}_v$ and $\mathcal{C}_l$ denote visual and language token indices, respectively, and $\mathbf{d}_{ij}$ is the cosine distance between output tokens. $\mathbf{s}_v$ and $\mathbf{s}_l$ are intra-modality diversities of visual and language modalities, respectively, while $\mathbf{s}_{vl}$ is inter-modality diversity.

Figure 3 illustrates these diversities across projection layer types and MLLM blocks. We observe three key properties of output token distributions: (1) Comparing $\mathbf{s}_v$, $\mathbf{s}_l$ and $\mathbf{s}_{vl}$ reveals notable difference across modalities. This confirms the need to compute intra- and inter-modality diversities separately to capture unique patterns of each modality; (2) both intra- and inter-modality diversities vary significantly across layer types. As shown in Figure 3 Top, value projection layers (v-proj) exhibit higher diversities than query (q-proj) and key (k-proj) projection layers, despite receiving identical input tokens within the same block, suggesting that value layers encode richer multimodal information; (3) these diversities fluctuate significantly across blocks, indicating that the capability to encode multimodal information varies with model depth.

To address these observations in MLLM pruning, we propose a layer-wise sparsity strategy based on multimodal output token diversity. Our core intuition is that layers with higher multimodal output token diversity should retain more parameters during pruning to maintain their capability to encode richer multimodal output tokens. We quantify the importance of each MLLM layer as the average of intra- and inter-modality diversities: $\mathbf{s} = (\mathbf{s}_v + \mathbf{s}_l + \mathbf{s}_{vl})/3$. Following ECoFLAP (Sung et al., 2024), sparsity is set inversely proportional to the precomputed layer's importance, ensuring that layers with higher diversities retain more parameters to preserve their representational capabilities.

3.3 Influence of Multimodal Input Tokens

We now shift our focus to the impact of input tokens in MLLM pruning. Our primary assumption is that input tokens from different modalities contribute distinctively to multimodal information processing. To investigate this, we first analyze attention distributions across blocks for each modality by computing the average attention scores per modality from the attention score matrix $\mathbf{A}$, as depicted in Figure 4. The result reveals a clear trend: different blocks put varying degrees of reliance on visual and language inputs. This variation in modality reliance across blocks implies that a static, uniform input activation calculation may be suboptimal.

This phenomenon motivates us to examine whether modality-specific input tokens contribute distinctively to pruning outcomes. To explore this, we conduct preliminary experiments comparing two approaches for computing input activations: (1) the conventional approach of using both visual and language input tokens ($||\mathbf{X}||_2$, denoted as "V+L"), and (2) a variant that focuses solely on language tokens ($||\mathbf{X}^l||_2$, denoted as "L"). We apply both methods across the 32 blocks of LLaVA-NeXT (Li et al., 2024a) to assess the influence of multimodal tokens at different network depths.

Table 1 presents the pruning results for the conventional approach (V+L) alongside the variant (L). Notably, including both visual and language tokens across all blocks achieves better performance in a visually rich information understanding task (e.g., ChartQA). In contrast, omitting visual tokens reduces performance on ChartQA but improves results on a multimodal understanding task (e.g., MME). These results confirm our intuition that different blocks engage with specific modalities to varying degrees for multimodal information processing, which influences pruning outcomes.

3.4 Adaptive Multimodal Input Activation

The above findings support the need for a pruning strategy that dynamically adapts to modality-specific contributions of individual blocks. To address this, we propose an adaptive method that selects multimodal input tokens for input activation calculations tailored to address each block's unique multimodal processing needs.

A key step in our approach is identifying core input tokens by measuring their contributions. In this work, we use the last row of the attention score matrix $\mathbf{A}$ as token contributions: $\mathbf{a} = \mathbf{A}[:, -1] \in \mathbb{R}^{N_v+N_l}$, which captures importance of multimodal tokens. This can guide the dynamic selection of input tokens based on the unique processing demand of each block. For example, in a layer emphasizing visual information, visual tokens have high contributions $\mathbf{a}$. Thus, more visual tokens are prioritized during the selection for that layer, ensuring that its input activation retains crucial visual features.

For input token selection, we adopt the data selection algorithm in Maharana et al. (2023) to select core input tokens while considering token diversity. This selection process prioritizes input tokens with high $\mathbf{a}$ values while ensuring that the output tokens they produce remain diverse in output token space. Specifically, we first update $\mathbf{a}$ by incorporating both the intrinsic and neighboring token contributions:

$$\mathbf{a}_i \leftarrow \mathbf{a}_i + \sum_{j \in \mathcal{N}_i} \mathbf{e}_{ij} \cdot \mathbf{a}_j, \quad \mathbf{e}_{ij} = \exp\left(-\gamma * \mathbf{d}_{ij}\right), \quad (3)$$

where $\mathcal{N}_i$ denotes ith token's nearest neighbors. We consider three nearest neighbors and $\gamma = 1$, following the default setting of the algorithm.

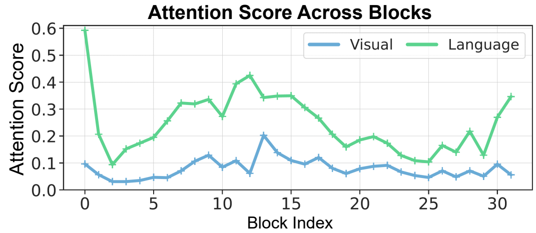


Figure 4: Average attention score across LLaVA-NexT blocks. Varying attention scores indicate that unique multimodal processing demands exist for each block.

Method	MME-cognition	MME-perception	ChartQA
Full Model	376.8	1588.3	69.2
V+L (Block 1 - 32)	276.4	1360.6	**63.2**
V+L (Block 1 - 2), L (Block 3 - 32)	**320.4**	**1476.6**	62.3
L (Block 1 - 32)	311.1	1468.4	62.2

Table 1: Impact of token selection on 50% pruning of LLaVA-NeXT across evaluation benchmarks. MME measures general multimodal understanding, while ChartQA focuses on visually rich information understanding (e.g., OCR, chart).

Once updated, we iteratively select the token with the highest contribution. To encourage selection diversity, the contributions of neighboring tokens of the selected token are penalized:

$$\mathbf{a}_j \leftarrow \mathbf{a}_j - \mathbf{e}_{ij} \cdot \mathbf{a}_i, \ \forall j \in \mathcal{N}_i, \quad (4)$$

where $\gamma = 0.2$, following the original setting. These iterative processes prioritize core multimodal tokens while minimizing redundant selections.

We select tokens from the full token index set $\mathcal{C}$ until the selected index set $\mathcal{C}'$ sufficiently represents the original distribution using maximum mean discrepancy (MMD) metric (Kim et al., 2016):

$$\text{MMD} = A(\mathcal{C},\mathcal{C}) + A(\mathcal{C}',\mathcal{C}') - 2A(\mathcal{C},\mathcal{C}') < 0.1 * \sqrt{\mathbf{s}},$$
$$A(\mathcal{C},\mathcal{C}') = \frac{1}{|\mathcal{C}||\mathcal{C}'|} \sum_{i \in \mathcal{C},\, j \in \mathcal{C}} \mathbf{e}_{ij}, \quad (5)$$

where $A(\mathcal{C},\mathcal{C}')$ measures the distributional similarity between two sets. The selection process continues until MMD falls below a threshold scaled by the function of $\mathbf{s}$, which accounts for variations in output token spaces, as shown in Figure 3. Input activation is then computed using the selected tokens: $||\mathbf{X}^{\mathcal{C}'}||_2$. This approach adaptively captures multimodal processing demands across different layers, which can facilitate pruning decisions by preserving parameters critical to those demands.

4 Experiments

4.1 Experimental Setups

Multimodal Large Language Models We conduct pruning on two popular MLLM architectures. LLaVA-NeXT (Li et al., 2024a) with 8B parameters enhances visual perception by splitting high-resolution images into sub-images. VideoLLaMA2 (Cheng et al., 2024) with 7B parameters improves spatiotemporal modeling and audio processing, making it well-suited for video and audio tasks. These models enable comprehensive evaluation of pruning strategies across diverse multimodal settings. A recent study (He et al., 2024) shows that pruning only the LLM component in MLLMs achieves a better balance between performance and efficiency since LLMs are typically much larger than these encoders. Therefore, our experiments focus on pruning the LLM component of MLLMs.

Evaluation Benchmark To assess performance after pruning, we evaluate their zero-shot capability on various multimodal benchmarks. We follow the evaluation protocols outlined in LLaVA-NeXT and VideoLLaMA2 to ensure consistent benchmark selection. For LLaVA-NeXT, we evaluate its zero-shot performance on multiple vision-language tasks: 1) multimodal understanding: MME (Fu et al., 2023) and MMMU (Yue et al., 2024); 2) visual mathematic reasoning: MathVista (Lu et al., 2024); 3) structural reasoning: ChartQA (Masry et al., 2022) and AI2D (Kembhavi et al., 2016); 4) multimodal perception: MMBench (Liu et al., 2024). For VideoLLaMA2, we assess its performance across diverse multimodal settings: 1) audio: Clotho-AQA (Lipping et al., 2022) for open-ended QA, TUT2017 (Mesaros et al., 2016) and VocalSound (Gong et al., 2022) for multiple-choice QA, and Muchomusic (Weck et al., 2024) for music understanding; 2) video: VideoMME and NeXTQA-MC for diverse video domains and durations, EgoSchema for long video understanding, and MVBench for spatio-temporal understanding; 3) audiovisual comprehension: MUSIC-QA (Li et al., 2022) for open-ended musical scene understanding. Further details on the evaluation pipeline are provided in Appendix A.

We report performances at sparsity ratios where pruned models maintain reasonably high performance that enables meaningful comparisons with baselines. To ensure fair comparisons across benchmarks with different scales, we compute the average relative performance, denoted as Rel., which measures model generalization. We compute relative performance as: (pruned model performance / full-model performance) $\times$ 100%.

Baselines We compare our method with several widely used pruning approaches. Magnitude (Zhu and Gupta, 2017), a standard baseline, removes weights with the smallest absolute values. SparseGPT (Frantar and Alistarh, 2023) is a layer-wise pruning method that leverages Hessian-based approximations to preserve critical weights. Wanda (Sun et al., 2024b) computes a layer-wise importance score as the product of weight magnitudes and input activations. OWL (Yin et al., 2024) proposes an outlier-weighted sparsity strategy, adjusting pruning ratios per layer based on outlier prevalence. ECoFLaP (Sung et al., 2024) uses zeroth-order gradient calculations to estimate the global importance score of VLM layers and determines layer sparsity ratios based on this score.

4.2 Results and Discussion

TAMP outperforms baselines on LLaVA-NeXT. Table 2 reports performance of LLaVA-NeXT at a 50% sparsity ratio. Across 6 of 7 benchmarks, including MME, AI2D, MMMU, Mathvista, and MMBench, TAMP ranks either first or second. On average, TAMP surpasses the strongest baseline by 1.9 percent points (pp) in relative performance, demonstrating its strength in preserving key parameters essential for versatile visual comprehension.

Furthermore, Figure 5 presents the performance of LLaVA-NeXT across a range of sparsity levels. TAMP exhibits the best performance-sparsity tradeoff. In contrast, pruning baselines experience steep accuracy declines beyond 50% sparsity, whereas our adaptive approach shows superior retention of model ability in high sparsity regimes (e.g., 60% and 70%), highlighting its robustness.

TAMP effectively preserves diverse multimodal understanding. To further examine our approach, we evaluate VideoLLaMA2 at a 60% sparsity ratio, with results presented in Table 3. TAMP ranks the top position in nearly all audio and video tasks and a close second in the audiovisual benchmark, outperforming the second-best baseline by 1.2 pp in average relative performance. These results demonstrate that our approach effectively captures modality-specific contributions, validating its universality across multiple modalities and tasks. Additional experiments on LLaVA-OneVision (Li et al., 2024c), which handles interleaved image and video modalities, are provided in Appendix B.

As shown in Figure 6, TAMP consistently maintains strong performance across different sparsity levels in VideoLLaMA2. This further shows

Method	MME-cognition	MME-perception	ChartQA	AI2D	MMMU	Mathvista	MMBench	Rel. (%)
Full Model	376.8	1588.3	69.2	71.7	40.1	36.2	72.2	100
Magnitude	0	0	0	0	24.0	26.6	0	19.0
SparseGPT	328.6	1448.9	**65.5**	64.5	33.6	31.3	64.7	89.0
Wanda	276.4	1360.6	63.2	64.3	**36.2**	30.2	63.9	86.0
ECoFLaP	254.6	1429.5	**65.5**	**66.1**	35.1	30.7	66.2	86.9
OWL	274.3	1366.0	63.2	64.0	35.3	30.8	64.1	85.9
TAMP (Ours)	**341.0**	**1470.2**	64.7	65.0	35.7	**31.9**	**66.3**	**90.9**

Table 2: Comparison of pruning techniques on the LLaVA-NeXT model with 50% sparsity ratio and estimate performance on various multimodal evaluation benchmarks. The best and the second best results are in **bold** and underlined, respectively.

Method	Audio				Video				Audiovisual	Rel. (%)
	Clotho -AQA	TUT 2017	Vocal Sound	Mucho music	Video MME	Ego Schema	NextQA -MC	MV -Bench	MUSIC -QA	
Full Model	85.6	71.2	92.4	58.9	48.7	49.3	73.3	58.4	79.4	100
Magnitude	0	0	0	25.8	0	20.8	20.3	0	0	12.6
SparseGPT	83.9	64.1	91.9	48.8	35.7	42.6	61.8	54.2	70.6	88.5
Wanda	83.1	65.6	**92.1**	51.4	39.4	44.4	65.0	53.2	**73.3**	91.0
ECoFLaP	83.7	67.2	**92.1**	54.4	41.3	**46.8**	69.8	54.2	73.0	93.8
OWL	83.2	**70.5**	91.3	47.6	37.7	43.6	63.1	52.4	68.9	89.4
TAMP (Ours)	**84.2**	69.9	**92.1**	**55.9**	**42.5**	46.7	**70.9**	**54.8**	72.6	**95.0**

Table 3: Comparison of pruning techniques on the VideoLLaMA2 model with 60% sparsity ratio and estimate performance on various multimodal evaluation benchmarks. The best and the second best results are in **bold** and underlined, respectively.

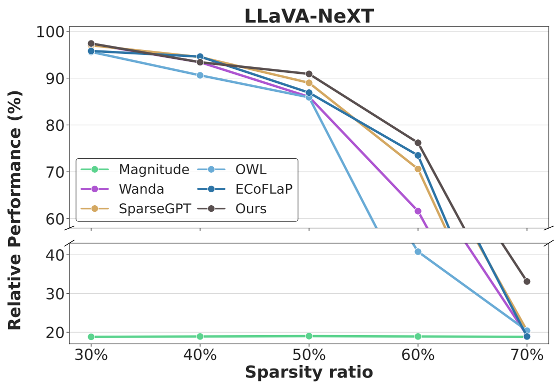


Figure 5: Average relative performances of all pruning techniques at different sparsity ratios for the LLaVA-NeXT.

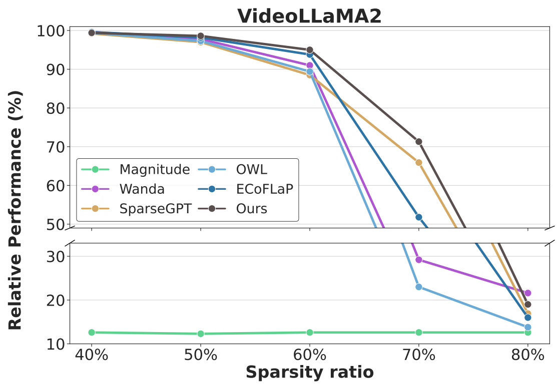


Figure 6: Average relative performances of all pruning techniques at different sparsity ratios for the VideoLLaMA2.

TAMP's robustness in maintaining diverse multimodal comprehension even under aggressive sparsity constraints. Moreover, in both Figure 5 and Figure 6, OWL suffers from severe performance drops at high sparsity ratios, unlike ECoFLaP and TAMP. OWL assigns layer-wise sparsity ratios proportional to the prevalence of outlier values within input activations computed across all input tokens. We hypothesize that multimodal encoder's tokens in MLLMs follow different outlier distributions than unimodal language tokens in LLMs, where large activation values in tokens are typically featured in LLMs (Sun et al., 2024a; Yin et al., 2024). This discrepancy likely contributes to OWL's underperformance, highlighting the importance of pruning strategies tailored for MLLMs to account for their unique multimodal attributes.

4.3 Further Analysis and Ablation

Core components in TAMP contribute to improving performance. To validate our strategies, we compare the two core components of TAMP, Diversity-Aware Sparsity (DAS) and Adaptive Multimodal Input Activation (AMIA), against ECoFLaP, OWL, and Wanda. Like DAS, ECoFLaP and OWL assign varying sparsity ratios to layers. AMIA selects core multimodal input tokens for input activations, while Wanda uses all input tokens. All the above methods build upon Wanda.

As shown in Table 4 (a), both DAS and AMIA bring substantial performance gain. DAS alone surpasses ECoFLaP and OWL, improving Wanda

(a) Key Components

Method	DAS	AMIA	LLaVA -NeXT	Video LLaMA2
Wanda	–	–	86.0	91.0
ECoFLaP	–	–	86.9	93.8
OWL	–	–	85.9	89.4
TAMP (Ours)	✓	–	<u>90.4</u>	<u>94.0</u>
	–	✓	89.5	92.3
	✓	✓	**90.9**	**95.0**

(b) Layer-wise Sparsity

Method	LLaVA -NeXT	Video LLaMA2
Wanda	86.0	91.0
+ All-token DAS	89.3	94.1
+ Block-wise DAS	88.0	94.2
+ DAS (Ours)	90.4	94.0
SparseGPT	89.0	88.5
+ DAS (Ours)	89.1	94.0

(c) Input Activation

Method	LLaVA -NeXT	Video LLaMA2
Wanda	86.0	91.0
Random	85.3	91.1
Attention	89.2	93.5
AMIA (Ours)	89.5	92.3

Table 4: Ablation studies of TAMP. DAS: Diversity-Aware Sparsity in Section 3.2, AMIA: Adaptive Multimodal Input Activation in Section 3.4. (a) Contributions of proposed components. (b) Ablation on layer-wise sparsity strategies. (c) The performance of different multimodal input token selections for input activations calculation. For all experiments, we prune LLaVA-NeXT at 50% and VideoLLaMA2 at 60% sparsity ratios, and report the relative average performance.

by 4.4 pp in LLaAV-NeXT and 3.0 pp in VideoLLaMA2. This supports the importance of multimodal token diversity in identifying critical layers for encoding rich multimodal representation. Notably, DAS outperforms ECoFLaP, which relies on gradient computations, despite calculating simple cosine distances among tokens. This demonstrates that DAS efficiently captures the complexities of multimodal data, further validating its efficacy.

AMIA improves Wanda by 3.5 pp in LLaVA-NeXT and 1.3 pp in VideoLLaMA2. This improvement stems from AMIA's adaptive selection of core multimodal tokens for input activations, aligning pruning with each layer's processing needs. Integrating DAS and AMIA, TAMP achieves superior performance, underscoring the advantage of jointly optimizing layer-wise sparsity and pruning decisions in MLLMs through multimodal attributes.

Ablation on layer-wise sparsity. We further test variants of DAS. All-token DAS averages the cosine distances of all output tokens to determine layer importance: $\mathbf{s} = \mathbb{E}_{i,j\sim\mathcal{C}}[\mathbf{d}_{ij}]$. Block-wise DAS averages the layer importance in DAS within each block and applies uniform sparsity to all layers in that block. The results are summarized in Table 4 (b). DAS shows robust performance across MLLMs compared to All-token DAS, validating the use of intra- and inter-modality diversities to reflect unique token distributions across modalities. DAS also improves SparseGPT's performance, further demonstrating its adaptability across pruning methods. Block-wise DAS outperforms ECoFLaP and OWL, demonstrating that our approach can also represent block-level importance. To provide deeper insights, we further analyze the sparsity ratios of baselines and our approach in Appendix C.

Adaptive Input Activation. To further validate our intuition that MLLM pruning needs to adapt to modality-specific contributions within each block, we conduct ablation studies on different token selection strategies for input activations. Our approach, AMIA, selects core tokens based on token contribution score $\mathbf{a}$ and output token distances. We examine other selection strategies: (1) Random, which randomly selects 100 tokens, and (2) Attention, which selects tokens with above-average contribution scores. As shown in Table 4 (c), attention-based selection methods, Attention and AMIA, outperform random selection, supporting our core intuition. However, the best strategy depends on the target MLLM. This may be due to the complexity of multimodal feature spaces, suggesting that further refinements in selection methods could enhance pruning robustness. We analyze AMIA's token selection with visualizations in Appendix D.

5 Conclusion

In this paper, we investigate the critical challenges in Multimodal Large Language Model (MLLM) pruning. In our comprehensive investigations, we find that different MLLM layers have varying capabilities to encode multimodal output tokens. We also empirically observe that existing pruning methods fail to address varying modality reliance across blocks in MLLMs, resulting in suboptimal pruning outcomes. Based on these observations, we introduce TAMP, a novel pruning framework that adapts both layer-wise sparsity and input activations to each layer's multimodal token attributes. We validate our method on powerful MLLMs and extensive experiments demonstrate that TAMP is effective in preserving diverse multimodal abilities, even at extreme model sparsity. We believe our work offers a strong foundation for future advancements in MLLM pruning, enabling the deployment of recent MLLMs in source-constrained scenarios.

Limitations

While TAMP shows promising performance on recent MLLMs, including LLaVA-NeXT, VideoLLaMA2, and LLaVA-OneVision, this work focuses on unstructured pruning for MLLMs. However, both the main body and the appendix reveal results consistent with recent structured pruning methods, indicating the potential of our core intuitions. Future research will investigate how our approach can be extended in this direction to enhance the applicability and efficiency of MLLM pruning techniques.

Despite evaluating TAMP on several MLLMs, MLLMs handling other modalities, such as point clouds, molecules, and proteins, or MLLMs incorporating Q-Former structures remain unexplored. Evaluating our approach across a border range of settings would further validate its generalizability. Additionally, our study primarily examines the performance-sparsity trade-offs without evaluating the impact on hardware efficiency. While unstructured pruning can theoretically reduce computation, future research in structured pruning should explore how TAMP can deliver practical benefits in terms of latency and deployment efficiency.

References

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

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024b. Sharegpt4v: Improving large multi-modal models with better captions. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. 2024. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Elias Frantar and Dan Alistarh. 2023. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. GPTQ: accurate post-training quantization for generative pre-trained transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

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. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, abs/2306.13394.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. BLINK: multimodal large language models can see but not perceive. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Yuan Gong, Jin Yu, and James R. Glass. 2022. Vocalsound: A dataset for improving human vocal sounds recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*.

Jinyang Guo, Jianyu Wu, Zining Wang, Jiaheng Liu, Ge Yang, Yifu Ding, Ruihao Gong, Haotong Qin, and Xianglong Liu. 2024. Compressing large language models by joint sparsification and quantization. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Shwai He, Ang Li, and Tianlong Chen. 2024. Rethinking pruning for vision-language models: Strategies for effective sparsity and performance restoration. *arXiv preprint arXiv:2404.02424*.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. MANTIS: interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Min Joon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Been Kim, Oluwasanmi Koyejo, and Rajiv Khanna. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024a. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Bo Li, Peiyuan Zhang, Kaichen Zhang, Fanyi Pu, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, et al. 2024b. Lmms-eval: Accelerating the development of large multimoal models.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024c. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024d. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2024. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10):1–42.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. AWQ: activation-aware weight quantization for on-device LLM compression and acceleration. In *Proceedings of the Seventh Annual Conference on Machine Learning and Systems, MLSys 2024, Santa Clara, CA, USA, May 13-16, 2024*.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. In *European Signal Processing Conference, (EUSIPCO)*.

Xuejing Liu, Wei Tang, Xinzhe Ni, Jinghui Lu, Rui Zhao, Zechao Li, and Fei Tan. 2023. What large language models bring to text-rich vqa? *arXiv preprint arXiv:2311.07306*.

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. 2024. Mmbench: Is your multi-modal model an all-around player? In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. Llm-pruner: On the structural pruning of large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Adyasha Maharana, Prateek Yadav, and Mohit Bansal. 2023. D2 pruning: Message passing for balancing diversity and difficulty in data pruning. *arXiv preprint arXiv:2310.07931*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics (ACL)*.

Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. 2024. Shortgpt: Layers in large language models are more redundant than you expect. *arXiv preprint arXiv:2403.03853*.

Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. 2016. TUT database for acoustic scene classification and sound event detection. In *European Signal Processing Conference (EUSIPCO)*.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*.

Oliver Sieberling, Denis Kuznedelev, Eldar Kurtic, and Dan Alistarh. 2024. Evopress: Towards optimal dynamic model compression via evolutionary search. *arXiv preprint arXiv:2410.14649*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. 2024a. Massive activations in large language models. *arXiv preprint arXiv:2402.17762*.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. 2024b. A simple and effective pruning approach for large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yi-Lin Sung, Jaehong Yoon, and Mohit Bansal. 2024. Ecoflap: Efficient coarse-to-fine layer-wise pruning for vision-language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Changyuan Wang, Ziwei Wang, Xiuwei Xu, Yansong Tang, Jie Zhou, and Jiwen Lu. 2024a. Q-vlm: Post-training quantization for large vision-language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024b. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*.

Benno Weck, Ilaria Manco, Emmanouil Benetos, Elio Quinton, George Fazekas, and Dmitry Bogdanov. 2024. Muchomusic: Evaluating music understanding in multimodal audio-language models. *arXiv preprint arXiv:2408.01337*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal LLM. In *Proceedings of the International Conference on Machine Learning (ICML)*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize

Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Zhewei Yao, Reza Yazdani Aminabadi, Minjia Zhang, Xiaoxia Wu, Conglong Li, and Yuxiong He. 2022. Zeroquant: Efficient and affordable post-training quantization for large-scale transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Lu Yin, You Wu, Zhenyu Zhang, Cheng-Yu Hsieh, Yaqing Wang, Yiling Jia, Gen Li, Ajay Kumar Jaiswal, Mykola Pechenizkiy, Yi Liang, Michael Bendersky, Zhangyang Wang, and Shiwei Liu. 2024. Outlier weighed layerwise sparsity (OWL): A missing secret sauce for pruning llms to high sparsity. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Lu Yu and Wei Xiang. 2023. X-pruner: explainable pruning for vision transformers. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, Hang Yan, Jie Fu, Tao Gui, Tianxiang Sun, Yu-Gang Jiang, and Xipeng Qiu. 2024. Anygpt: Unified multimodal LLM with discrete sequence modeling. In *Proceedings of the Association for Computational Linguistics (ACL)*.

Mingyang Zhang, Hao Chen, Chunhua Shen, Zhen Yang, Linlin Ou, Xinyi Yu, and Bohan Zhuang. 2024a. Loraprune: Structured pruning meets low-rank parameter-efficient fine-tuning. In *Findings of the Association for Computational Linguistics (ACL)*.

Yang Zhang, Yawei Li, Xinpeng Wang, Qianli Shen, Barbara Plank, Bernd Bischl, Mina Rezaei, and Kenji Kawaguchi. 2024b. Finercut: Finer-grained interpretable layer pruning for large language models. *arXiv preprint arXiv:2405.18218*.

Yingtao Zhang, Haoli Bai, Haokun Lin, Jialin Zhao, Lu Hou, and Carlo Vittorio Cannistraci. 2024c. Plug-and-play: An efficient post-training pruning method for large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Longguang Zhong, Fanqi Wan, Ruijun Chen, Xiaojun Quan, and Liangzhi Li. 2024. Blockpruner: Fine-grained pruning for large language models. *arXiv preprint arXiv:2406.10594*.

Aojun Zhou, Yukun Ma, Junnan Zhu, Jianbo Liu, Zhijie Zhang, Kun Yuan, Wenxiu Sun, and Hongsheng Li. 2021. Learning N: M fine-grained structured sparse neural networks from scratch. *arXiv preprint arXiv:2102.04010*, abs/2102.04010.

Michael Zhu and Suyog Gupta. 2017. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*.

A Details of Experimental Setups

Calibration Datasets. Following established practices in model pruning (Sun et al., 2024b; Frantar and Alistarh, 2023; Sung et al., 2024), we use a random subset of 128 samples from the training datasets of the target models as calibration data. For LLaVA-NeXT, we use ShareGPT4V (Chen et al., 2024b) as the calibration dataset. For VideoLLaMA2, we choose MUSIC-QA (Li et al., 2022) as the calibration source as its samples consist of both video and audio modalities. For LLaVA-OneVision, we use NLVR2 (Suhr et al., 2019) as it constitutes the largest portion of its training dataset.

Evaluation pipeline To ensure consistency and reproducibility, the benchmarks are assessed through LMMs-Eval framework (Li et al., 2024b) and evaluation pipelines of the models. We follow the LMMs-Eval prompt templates provided in the official GitHub repository of the LMMs-Eval to evaluate the LLaVA-NeXT and LLaVA-OneVision models. We implement the VideoLLaMA2 architecture on the LLMs-Eval framework and evaluate the model on audio and video benchmarks.

B Experiments on LLaVA-OneVision

Experimental Setups We conduct additional model pruning experiments on LLaVA-OneVision (Li et al., 2024c) with 7B parameters, which processes both interleaved images and video modalities. After pruning, we evaluate its zero-shot performance in the two modalities, following the evaluation protocols in LLaVA-OneVision: 1) interleaved images: Muirbench (Wang et al., 2024b) for diverse multi-image tasks, Mantis (Jiang et al., 2024) for reasoning over multiple images, BLINK (Fu et al., 2024) for multi-image visual perception tasks, and Text-rich VQA (Liu et al., 2023) for multi-image text recognition; 2) video: VideoMME and NeXTQA-MC for diverse video domains and durations, EgoSchema for long video understanding, and MVBench for spatio-temporal understanding.

Experimental Results Table 5 summarizes performance of LLaVA-OneVision at a 50% sparsity ratio. Across 6 out of 8 interleaved images understanding and video benchmarks, TAMP ranks either first or second. On average, TAMP surpasses the Wanda and the strongest baseline by 5.3 pp and 0.3 pp, respectively, in relative performance. These results demonstrate that the effectiveness of our method can be transferred to the pruning of other recent MLLMs with different multimodal settings, further supporting the universality of our approach.

C In-Depth Analysis on Layer-wise Sparsity Ratios

C.1 Sparsity of Projection Layer Type

In Figure 3, we observe significant variations in intra- and inter-modality diversities across different projection layer types and leverage these variations to estimate layer importance. In this ablation study, we examine the sparsity results of different projection layer types in MLLMs. Figure 7 presents the average sparsity ratios across blocks for each projection layer type in VideoLLaMA2 pruned at 70% sparsity using TAMP. Our analysis reveals that in the MHA module, comprising query, key, value, and output projection layers, the value projection layer consistently exhibits the lowest sparsity ratio. In contrast, in the FFN module, which consists of gate, up, and down projection layers, all projection layers exhibit relatively high sparsity levels compared to the layers in the MHA module, with the gate projection layer showing the highest value.

These findings suggest that FFN modules are more robust in pruning than MHA modules, which aligns with the recent work (Zhang et al., 2024b) on pruning either MHA or FFN modules in LLMs. Moreover, our results imply that value projection layers may play a more crucial role in encoding token features compared to other projection layers, containing more critical parameters necessary for preserving the performance of MLLMs.

Interestingly, the aforementioned trend in average sparsity per layer type shown in Figure 7 is consistent with the result in EvoPress (Sieberling et al., 2024), which uses an evolutionary algorithm to find the optimal sparsity levels across LLM layers or blocks. This alignment supports our core intuition that layers producing higher multimodal output token diversity should retrain more parameters during pruning to preserve their capability to encode richer multimodal information. Moreover, TAMP uncovers this trend through a simpler approach based on average cosine distances among multimodal tokens, while EvoPress requires iterative exploration in the vast sparsity ratio solution space. This indicates that output token distributions can offer an efficient and insightful basis for estimating layer importance, leading to more effective pruning outcomes.

Method	Interleaved Image Muir Bench	Mantis	BLINK	Text-rich VQA	Video Video MME	Ego Schema	NextQA -MC	MV -Bench	Rel. (%)
Full Model	41.8	64.2	48.4	80.1	60.1	56.7	58.4	79.4	100
Magnitude	0	0	0	0	20.7	0	0	19.8	7.4
SparseGPT	**41.1**	51.9	**46.2**	63.6	**55.2**	**55.1**	52.3	75.2	90.9
Wanda	38.6	47.9	44.0	65.1	51.5	53.5	49.3	73.3	87.0
ECoFLaP	40.7	**58.1**	45.2	64.7	54.4	54.5	53.2	**76.1**	92.0
OWL	35.1	43.8	43.5	60.1	50.8	52.4	47.6	69.1	82.8
TAMP (Ours)	40.9	57.1	45.9	**69.8**	54.0	53.9	52.5	75.2	**92.3**

Table 5: Comparison of pruning techniques on the LLaVA-OneVision model with 60% sparsity ratio and estimate performance on various multimodal evaluation benchmarks. The best and the second best results are in **bold** and underlined, respectively.

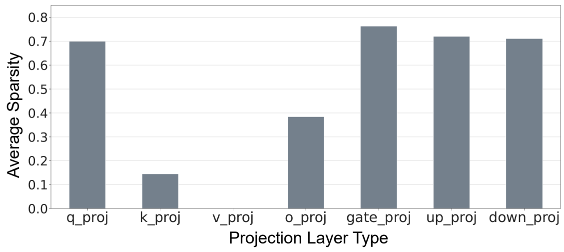


Figure 7: Average sparsity per projection layer type for VideoLLaMA2 at 70% sparsity using TAMP (Ours).

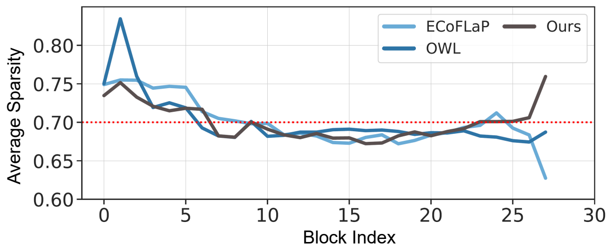


Figure 8: Comparison of sparsity ratio results per block for VideoLLaMA2 model at 70% sparsity.

C.2 Block-wise Average Sparsity

To further investigate layer-wise sparsity strategies, we analyze sparsity ratios across block depths for VideoLLaMA2 at 70% sparsity ratio, as determined by ECoFLaP, OWL, and TAMP. The results illustrated in Figure 8 indicate that all three methods follow a similar sparsity trend, where the initial blocks have high sparsity and intermediate blocks exhibit moderate sparsity. Notably, TAMP exhibits higher sparsity in the last blocks.

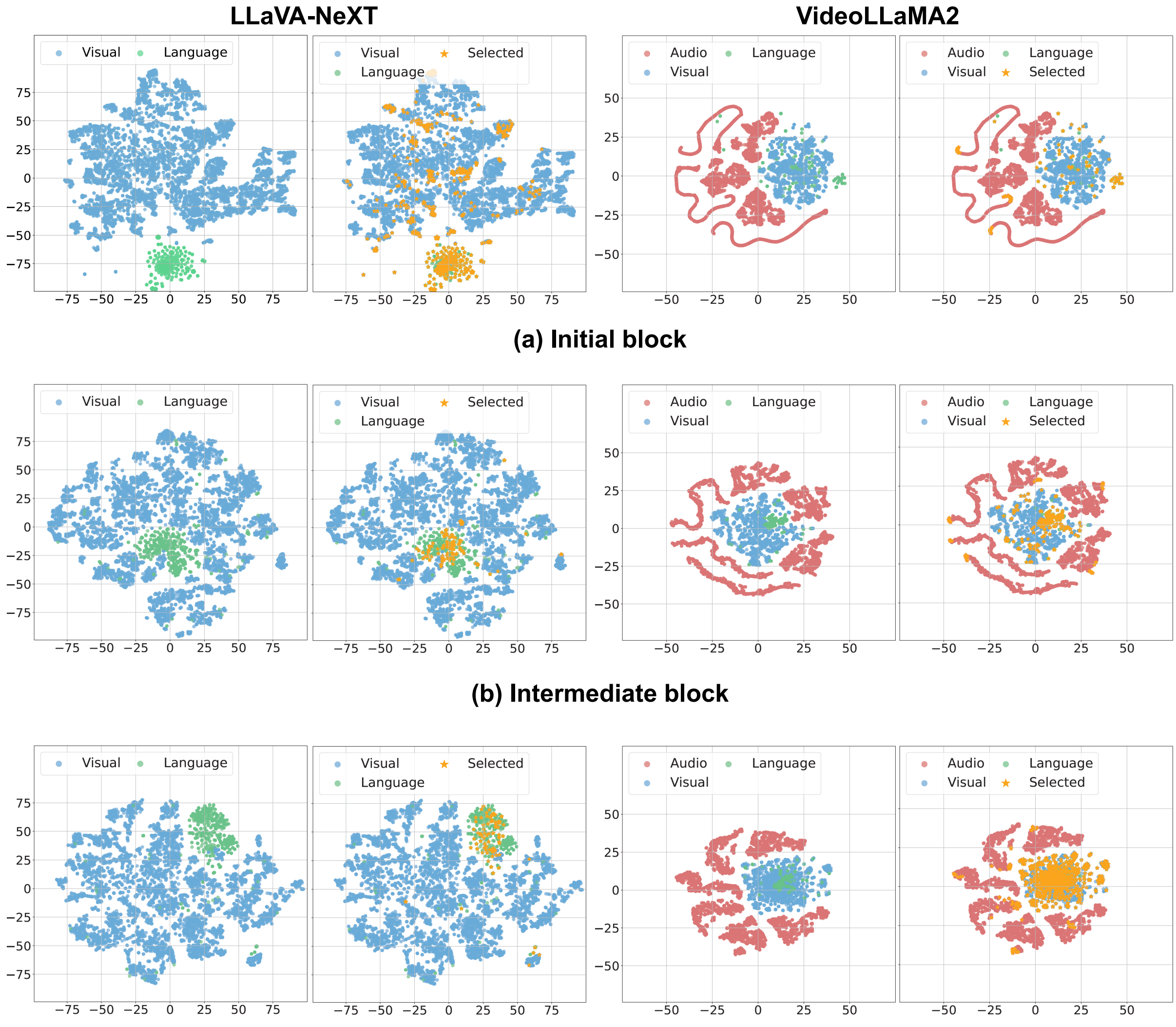
These trends diverge from typical observations in LLM pruning. Recent studies suggest that intermediate blocks generally contain large redundancy, where pruning these blocks results in a mere impact on LLM performance. In contrast, the studies show that pruning early or final blocks leads to substantial performance degradation (Men et al., 2024; Zhong et al., 2024). However, in MLLM pruning in Figure 8, we observe an opposite pattern, particularly with TAMP.

This difference can be explained by the attention distribution trends shown in Figure 4. Our analysis reveals that both visual and language attention scores are notably high in the intermediate blocks, indicating active multimodal interactions. Thus, these blocks would require lower sparsity to align with the increased multimodal integration occurring at this stage. In contrast, in the later blocks, only their language attention scores are high while visual attention scores are low. We hypothesize that at this stage, the MLLM primarily focuses on language generation, reducing the need for multimodal processing, which aligns with token reduction studies in MLLMs (Chen et al., 2024a). This suggests that non-language modality information becomes redundant in these layers, requiring comparably fewer parameters.

These findings support the necessity of pruning strategies specifically designed for MLLMs, as their architectural and functional characteristics differ significantly from those of LLMs. Conventional LLM pruning techniques may therefore be suboptimal for multimodal models.

D Visualizing Multimodal Selection

In Figure 9, we illustrate token selection results using the AMIA selection strategy across the initial, intermediate, and last blocks of LLaVA-NeXT and VideoLLaMA2. Specifically, we visualize the multimodal output token spaces from the value projection layers in each block. Each block exhibits different modality distributions and selection results, showcasing the varying multimodal processing demands across different blocks. For example, in LLaVA-NeXT, the initial block selects both visual and language tokens, indicating its need for



(a) Initial block

(b) Intermediate block

(c) Last block

Figure 9: Token selection results of Adaptive Multimodal Input Activation. We use t-SNE visualization for multimodal output token space across value projection layers of initial, intermediate, and last blocks of LLaVA-NeXT and VideoLLaMA2.

visual information during multimodal processing. However, in the intermediate and final blocks of the LLaVA-NeXT, comparably fewer visual tokens are selected, suggesting that these blocks assign less importance to visual information compared to language information. In VideoLLaMA2, the last block continues to rely on both language and visual tokens. We attribute this to VideoLLaMA2's architecture, which processes video inputs through spatial-temporal aggregation. This design yields more compact and informative video tokens compared to approaches that simply segment video into small image pacthes (Cheng et al., 2024; Li et al., 2024d).