When Language Overrules: Revealing Text Dominance in Multimodal Large Language Models

Huyu Wu[1] **Meng Tang**[2] **Xinhan Zheng**[1,3] **Duo Su**[4] **Haiyun Jiang**[5*]
[1]Beijing Angelpro Technology Co., Ltd. [2]Department of Computer Science, Aberystwyth University
[3]University of Science and Technology of China [4]Tsinghua University [5]Fudan University
huyu-wu@outlook.com haiyunjiangnlp@gmail.com

Abstract

Text dominance, the tendency of multimodal large language models (MLLMs) to over-attend to textual tokens while under-utilizing non-text inputs, has been observed in vision-language settings, but whether it extends to other modalities remains an open question. We present a systematic, attention-level study of this phenomenon across five modalities (image, video, audio, time-series, and graph), using two diagnostic metrics: the Modality Dominance Index (MDI), which compares per-token attention between text and non-text inputs, and the Attention Efficiency Index (AEI), which normalizes attention share by token share. Across ten model-benchmark pairs and five benchmarks, text dominance appears in the tested image, video, audio, and time-series settings and typically intensifies in deeper layers. Controlled token-replication experiments show that expanding non-text sequences without adding semantic content amplifies the imbalance, while graph tasks provide a boundary case where compact, information-dense tokens reverse the effect. Guided by these findings, we evaluate attention-based token compression as a proof-of-concept intervention on the vision modality. On LLaVA-1.5-7B with MMMU-Pro, removing 90% of visual tokens via informed selection preserves accuracy (33.70% vs. 33.23% baseline), while random dropping under the same ratio degrades it to 30.29%; informed selection also lowers the late-layer MDI from 15.63 to 3.46 and reduces latency by 3.5×. These results suggest that text dominance is a cross-modal phenomenon associated with token redundancy, and that token compression can reduce it in the tested vision setting while preserving task performance.

1 Introduction

Multimodal large language models (MLLMs) Yin et al. [2024], Qin et al. [2025], Team et al. [2025], Bai et al. [2025] have achieved strong results across vision, audio, and structured-data tasks by combining modality-specific encoders with a large language model backbone. Despite these advances, a growing body of evidence suggests that the integration is far from balanced: MLLMs tend to allocate disproportionate attention to textual tokens while under-utilizing non-text inputs Cai et al. [2025], Zheng et al. [2025], Jia et al. [2025], a pattern we call *text dominance*.

Evidence of this imbalance has so far been reported primarily in vision-language settings. For instance, VQA models can answer questions correctly even when the input image is removed Liu et al. [2024b], suggesting that text tokens alone carry sufficient signal for the model's prediction. Leng et al. [2024] proposed the Modality Importance Score for video question answering, extending the observation beyond static images. However, these studies are limited to one or two modality

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

pairs, leaving open the question of whether text dominance is specific to certain architectures or a more general property of current MLLMs.

In this work, we ask whether text dominance is merely an artifact of specific vision-language architectures, or a broader phenomenon that persists across modalities and network depths in current MLLMs.

To answer this question, we conduct a systematic attention-level analysis across five modalities: image, video, audio, time-series, and graph. We define two metrics: the Modality Dominance Index (MDI), which compares per-token attention between text and non-text inputs, and the Attention Efficiency Index (AEI), which normalizes each modality's attention share by its token share.

Our analysis yields three findings. First, text dominance is pervasive: across ten models, it appears in image, video, audio, and time-series settings and intensifies in deeper layers, with the late-layer MDI reaching 157.53 for VideoLLaMA3-7B on MMBench-Video Fang et al. [2024]. Second, controlled token-replication experiments show that expanding non-text sequences without adding semantic content consistently amplifies the imbalance. This supports the hypothesis that token redundancy drives attention dilution: as non-text tokens proliferate, per-token attention is spread thinner, favoring the more compact text modality. Third, graph tasks provide a boundary case: when non-text tokens are compact and information-dense, the dominant modality can reverse, suggesting that the phenomenon depends on the interaction between token count and information density.

Motivated by these findings, we evaluate attention-guided token compression Chen et al. [2024], Zhang et al. [2024] as a proof-of-concept intervention on the vision modality, where mature compression methods exist. On LLaVA-1.5-7B with MMMU-Pro Yue et al. [2024], removing 90% of visual tokens via informed selection preserves accuracy (33.70% vs. 33.23% baseline), while random dropping under the same ratio degrades it to 30.29%. The informed selection also reduces the late-layer MDI from 15.63 to 3.46, suggesting that many of the removed tokens were redundant in this setting.

To our knowledge, this is the first attention-level study of text dominance that spans five modalities and ten models under a unified diagnostic framework (MDI and AEI). Beyond characterizing the phenomenon, we provide controlled evidence that non-text token proliferation amplifies text dominance, and observational evidence that task formulation further modulates the effect. We also show, as a proof of concept on the vision modality, that existing token compression methods can reduce text dominance while preserving accuracy and lowering latency.

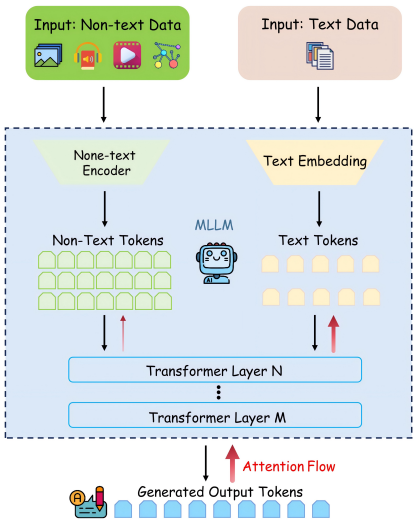

Figure 1: Overview of text dominance in MLLMs. Each modality is tokenized and fed to the LLM backbone. The thinner arrows from non-text tokens reflect their lower per-token attention: despite comprising the majority of input tokens, non-text modalities receive substantially less attention per token than text.

2 Related Work

2.1 Modality Imbalance: Evidence and Explanations

MLLMs typically combine a pretrained modality-specific encoder with a large language model backbone via a learned interface Liang et al. [2024]. This architecture has been extended to video Zhang et al. [2025], audio Chu et al. [2024], time-series Xie et al. [2024], and graph data Tang et al. [2024], yet a recurring observation across these settings is modality imbalance Prabhu [2025]: the model over-relies on text while under-utilizing non-text inputs.

The roots of this imbalance have been attributed to both data and architecture. On the data side, the higher information density of text relative to raw perceptual input creates exploitable shortcuts Park et al. [2025]. On the architecture side, most MLLMs pair an LLM pretrained on trillions of text tokens with a vision encoder trained on a comparatively smaller corpus Li et al. [2023], Liu et al. [2023], creating an asymmetry in backbone capacity. Prior work has studied this imbalance primarily in image-text settings, without attention-level quantification across modalities. We provide such a cross-modal analysis, covering five modalities and ten model-benchmark pairs with a unified diagnostic framework.

2.2 Token Efficiency and Compression in MLLMs

A complementary line of work seeks to mitigate modality imbalance through training-time or inference-time interventions. The Data Remixing framework Ma et al. [2025] masks the stronger modality during training to force reliance on the weaker one. The MBPO framework Liu et al. [2025] applies Direct Preference Optimization on adversarially generated negatives to reduce text-driven hallucination. At inference time, methods such as FastV Chen et al. [2024] and FasterVLM Zhang et al. [2024] compress visual tokens to improve efficiency. These approaches target efficiency or alignment but do not connect token compression to modality balance. We repurpose existing compression methods as controlled tools to test the attention dilution hypothesis, rather than proposing a new compression algorithm.

3 Text Dominance in Multimodal Large Language Models

To move beyond qualitative observations of modality bias, we require a rigorous quantitative framework. This section formalizes the measurement of text dominance: we first describe the benchmarks and models (§3.1), then define two diagnostic metrics (§3.2), and present the main results (§3.3).

3.1 Datasets and Models

We evaluate across five modalities, selecting benchmarks intended to reduce text-only shortcuts. For image, we use MMMU-Pro Yue et al. [2024], which filters out text-only-answerable questions, and test Qwen2.5-VL (7B/32B/72B) Bai et al. [2025], LLaVA-1.5-7B Liu et al. [2024a], and Kimi-VL-A3B-Instruct Team et al. [2025] to cover diverse model scales. For video, MMBench-Video Fang et al. [2024] provides long-form temporal reasoning tasks, evaluated on Qwen2.5-VL-7B and VideoLLaMA3-7B Zhang et al. [2025]. Audio, time-series, and graph are each represented by one domain-standard benchmark, IEMOCAP Busso et al. [2008], TS-Reasoning Xie et al. [2024], and GraphGPT-Eval Tang et al. [2024], paired with the corresponding specialist model (Qwen2-Audio-7B-Instruct Chu et al. [2024], ChatTS-14B, and GraphGPT-7B). Full details of all model-benchmark pairs are listed in Table 1.

3.2 Evaluation Metrics

To characterize how an MLLM allocates its computational resources across modalities, we employ two complementary indices: the *Modality Dominance Index* (MDI) and the *Attention Efficiency Index* (AEI). The MDI captures overall modality dominance in generation, whereas the AEI measures the attention efficiency of each modality relative to its token proportion.

Let $\mathcal{T}$ and $\mathcal{O}$ denote the sets of text tokens and non-text tokens, respectively. For the generation of N output tokens, let $a_{i,j}$ denote attention from output token i to input token j, averaged over selected

layers and all heads, then normalized over input tokens at each step. We define the total attention mass received by text and non-text tokens as

$$A_T = \sum_{i=1}^{N} \sum_{j \in \mathcal{T}} a_{i,j}, \qquad A_O = \sum_{i=1}^{N} \sum_{j \in \mathcal{O}} a_{i,j}. \tag{1}$$

Definition 3.1 (Modality Dominance Index)**.** *We define the Modality Dominance Index as*

$$\mathrm{MDI} = \frac{A_T/|\mathcal{T}|}{A_O/|\mathcal{O}|}. \tag{2}$$

The MDI compares the average attention received by one text token and one non-text token. A value larger than 1 indicates per-token text dominance, a value smaller than 1 indicates per-token non-text dominance, and a value close to 1 indicates similar average attention per token across the two modalities.

Definition 3.2 (Attention Efficiency Index)**.** *Let*

$$P_T = \frac{A_T}{A_T + A_O}, \qquad Q_T = \frac{|\mathcal{T}|}{|\mathcal{T}| + |\mathcal{O}|}, \tag{3}$$

where P_T is the attention share of text and Q_T is the token share of text. We define the Attention Efficiency Index for the text modality as

$$\mathrm{AEI} = \frac{P_T}{Q_T} = \frac{A_T/(A_T + A_O)}{|\mathcal{T}|/(|\mathcal{T}| + |\mathcal{O}|)}. \tag{4}$$

The AEI measures how effectively text converts its token budget into model attention. An AEI value larger than 1 indicates that text receives more attention than its token proportion would suggest, while a value smaller than 1 indicates that text is under-attended relative to its input size. Throughout this paper, AEI refers to the text modality unless otherwise specified.

MDI and AEI are complementary: MDI measures per-token dominance, while AEI normalizes attention by token budget; the two are related but not redundant. Both are diagnostic statistics of attention allocation, not causal measures of modality importance Jain and Wallace [2019]. We therefore report them jointly with accuracy and latency throughout the paper, and interpret shifts only under controlled interventions (§4) or matched baselines (§5).

Model	Modality	Dataset	Early		Middle		Late	
			MDI	AEI	MDI	AEI	MDI	AEI
Qwen2.5-VL-7B	Image	MMMU_Pro	2.26	1.42	21.12	10.86	33.10	14.24
Qwen2.5-VL-32B			3.84	2.82	54.96	21.88	26.03	13.95
Qwen2.5-VL-72B			9.33	6.15	92.21	60.43	24.46	14.60
LLaVA-1.5-7B			1.91	1.55	8.07	2.92	15.63	3.36
Kimi-VL-A3B-Instruct			2.27	3.91	3.78	2.99	28.39	2.59
Qwen2.5-VL-7B	Video	MMBench-Video	10.72	9.60	74.13	41.78	86.95	47.84
VideoLLaMA3-7B			19.14	17.90	140.10	73.75	157.53	76.26
Qwen2-Audio-7B-Instruct	Audio	IEMOCAP $\times$1	1.02	1.32	3.24	1.99	1.16	1.08
		IEMOCAP $\times$5	2.65	2.56	8.09	5.17	6.73	4.31
		IEMOCAP $\times$10	2.80	2.50	10.10	5.46	8.70	5.09
ChatTS-14B	Time-series	TS-Reasoning $\times$1	1.52	1.19	4.37	1.40	3.52	1.37
		TS-Reasoning $\times$5	2.08	1.95	10.72	3.15	9.28	3.03
		TS-Reasoning $\times$10	2.36	2.67	20.70	5.37	16.25	5.13
GraphGPT-7B	Graph	GraphGPT-Eval $\times$1	0.14	0.84	0.14	0.84	0.20	0.90
		GraphGPT-Eval $\times$5	0.20	0.69	0.35	0.83	0.69	0.98
		GraphGPT-Eval $\times$10	0.31	0.71	0.68	0.97	1.35	1.14

Table 1: **MDI and AEI across models, modalities, and benchmarks.** Image and video rows use the original input without replication. The notation $\times n$ denotes the replication factor applied to non-textual tokens. Early, Middle, and Late report statistics from the first two, middle two, and last two model layers, respectively. GraphGPT-Eval and TS-Reasoning abbreviate GraphGPT-Eval-Instruction and TimeSeries-Reasoning. Rows with replication factors ($\times$5, $\times$10) are analyzed in §4.

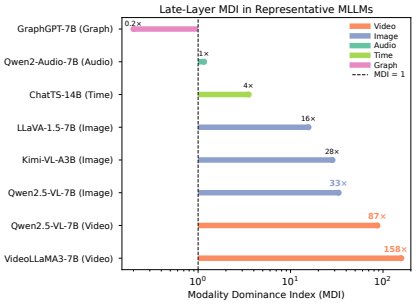

Figure 2: Text dominance across MLLMs of different modalities. The dashed line marks MDI = 1; points to its right indicate text-dominant attention. The graph modality (MDI < 1) is a boundary case discussed in §4.3, where compact, information-dense tokens reverse the dominant modality.

3.3 Experimental Results

Table 1 reports MDI and AEI across all model-benchmark pairs, aggregated over early, middle, and late layers; rows with replication factors ($\times 5$, $\times 10$) are analyzed in §4. The results reveal a clear and consistent pattern: text dominance is prevalent across most tested modalities, with graph as a notable exception, and typically intensifies in deeper layers.

Text dominance is pervasive in image and video tasks. For Qwen2.5-VL-7B on MMMU-Pro, the MDI rises from 2.26 in early layers to 33.10 in late layers, where image tokens receive 54.4% of total input attention. VideoLLaMA3-7B reaches a late-layer MDI of 157.53 on MMBench-Video, with video tokens receiving 50.9% of total input attention (Appendix E).

We also observe that text dominance varies substantially across modalities and appears to covary with non-text token count (Table 1); Section 4 investigates the drivers of this variation through controlled experiments.

Is text dominance aligned with task relevance? A natural question is whether text dominance reflects a rational allocation strategy, i.e., whether the model ignores non-text tokens precisely when they are uninformative. To test this, we sort MMMU-Pro samples by their MDI on LLaVA-1.5-7B and divide them into five equal-sized buckets, then compare accuracy under the original multimodal input (*With Image*) and a text-only ablation (*Text Only*) that removes all image tokens (Figure 3).

The results reveal a striking paradox: the accuracy gap Δ between *With Image* and *Text Only* shows an overall increase from -5 percentage points in the lowest MDI bucket to $+20$ in the highest. In other words, the samples where the model most heavily favors text tokens (MDI $\in [6.0, 15.6]$) are precisely the ones where images are most critical: removing images drops accuracy from 30% to 10%. Conversely, in the lowest MDI bucket (MDI $\in [0.7, 1.8]$), images appear unnecessary or even harmful (accuracy improves from 15% to 20% without them), yet the model allocates relatively more attention to image tokens in this group. This inverse relationship suggests that, at least in this setting, text dominance is not well-aligned with task relevance: the model under-attends to non-text tokens most severely when they carry the most decision-relevant information. A qualitative case study illustrating this failure mode on individual samples is provided in Appendix F.

Having established that text dominance is pervasive and that attention allocation appears misaligned with task relevance, a natural question follows: what input factors are associated with stronger text dominance? Section 4 investigates this through controlled and observational evidence.

4 What Drives Text Dominance? Empirical Evidence

The consistent escalation of MDI in deeper layers (Table 1) suggests that input characteristics shape text dominance during generation. We hypothesize that a major contributor is *attention dilution* stemming from token redundancy: non-text modalities are encoded into far more tokens than the accompanying text, yet many of these tokens are semantically redundant, causing attention to spread thinly across them. This section examines two factors that are consistently associated with stronger

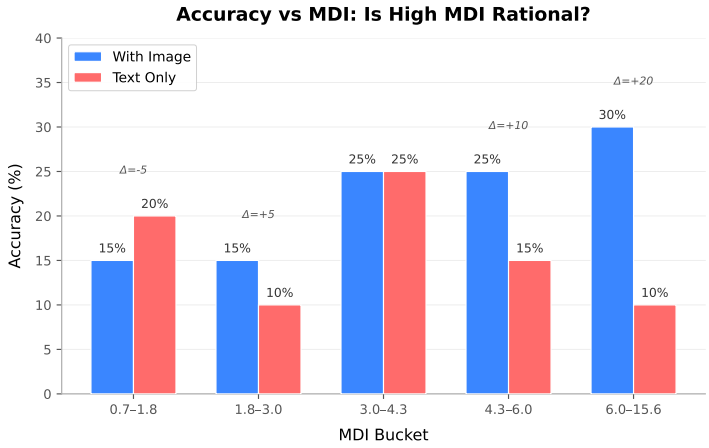

Figure 3: Accuracy under multimodal (*With Image*) and text-only (*Text Only*) inputs across MDI buckets on LLaVA-1.5-7B / MMMU-Pro. The accuracy gap Δ increases with MDI, indicating that the model under-attends to images most severely on samples where visual information is most critical for correct prediction.

text dominance in our data: (i) token redundancy in non-text modalities, tested through controlled replication experiments, and (ii) text-centric task formulations, examined through observational comparison. Graph tasks provide a boundary case that further supports the token-redundancy account (§4.3). For the first factor, we repeat the same non-text representations while varying token count, measuring the combined effect of sequence expansion and repeated content in the tested settings. For the second factor, we present observational evidence that is suggestive but not causally identified, since task formulation co-varies with other design choices.

4.1 Token Redundancy and Attention Dilution

Non-text modalities are typically encoded into far more tokens than the accompanying text prompt, yet many of these tokens carry overlapping information. If redundancy dilutes per-token attention, then expanding the non-text sequence without adding information should amplify text dominance.

We test this prediction with controlled token replication: we fix the text prompt and repeat the non-text token sequence by factors of $\times 1$, $\times 5$, and $\times 10$ (Table 1). Across three modalities, late-layer MDI rises monotonically: Qwen2-Audio-7B-Instruct ($1.16 \rightarrow 6.73 \rightarrow 8.70$), ChatTS-14B ($3.52 \rightarrow 9.28 \rightarrow 16.25$), and GraphGPT-7B ($0.20 \rightarrow 0.69 \rightarrow 1.35$). The last case is informative: a model that is initially non-text-dominant (MDI = 0.20) becomes text-dominant (MDI = 1.35) solely through token expansion, without any change in semantic content.

The video setting provides the most extreme natural illustration of this pattern. VideoLLaMA3-7B encodes hundreds of frame tokens per video, reaching a late-layer MDI of 157.53 on MMBench-Video, with a corresponding AEI of 76.26. Each text token receives substantially more attention on average, while video tokens collectively receive 50.9% of total input attention in the last two layers.

4.2 Task Formulation as an Additional Factor

Task formulation appears to be an additional factor associated with text dominance beyond token count alone. In time-series tasks, key normalization factors and task-specific metadata are encoded in natural language instructions, making the text prompt structurally central to inference. Audio emotion recognition tasks similarly define the reasoning scaffold through textual prompts.

Even at $\times 1$ replication (Figure 4), the textual modality already achieves AEI > 1 in late layers for both ChatTS-14B (1.37) and Qwen2-Audio-7B-Instruct (1.08). As non-text tokens are replicated, this advantage becomes more pronounced: for ChatTS-14B, the late-layer AEI increases from 1.37 ($\times 1$) to 3.03 ($\times 5$) and 5.13 ($\times 10$). A similar trend holds for Qwen2-Audio-7B-Instruct, where the late-layer AEI rises from 1.08 to 4.31 and 5.09 under the same conditions.

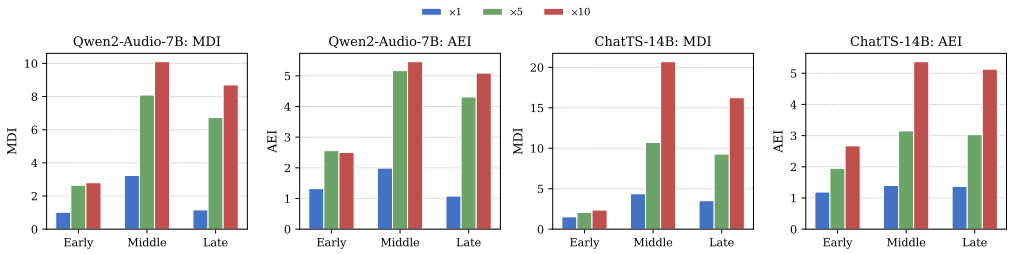

Figure 4: MDI and AEI of audio and time-series models under $\times 1$, $\times 5$, and $\times 10$ token replication. Even at $\times 1$, both models show AEI > 1 in late layers, suggesting that task formulation contributes to text dominance beyond token scaling alone.

We note that this observation is correlational: we cannot fully separate the influence of task formulation from other factors such as pretraining data composition. The key takeaway is that text dominance can be present even when non-text tokens are not heavily replicated, which suggests that token redundancy alone does not fully account for the phenomenon.

4.3 Graph Tasks: A Boundary Case

Our framework predicts that if non-text tokens are compact and information-dense, the dominance should reverse. Graph tasks are consistent with this prediction. Under standard conditions, GraphGPT-7B yields a late-layer MDI of 0.20, with textual AEI at 0.90, indicating that graph tokens attract more attention per token than text. When we inflate the graph token count by $\times 5$ and $\times 10$ replication without altering semantic content, MDI rises to 0.69 and 1.35, respectively, crossing the equilibrium threshold. The textual AEI shifts correspondingly from 0.90 to 0.98 and 1.14.

These observations collectively suggest that token count and information density are important determinants of modality dominance in the tested settings, consistent with the attention dilution hypothesis. Together with the token-scaling results above, these findings provide converging evidence that text dominance is associated with the expansion of semantically redundant non-text tokens. This observation naturally motivates our intervention strategy in the next section.

5 Can Token Compression Mitigate Text Dominance?

The attention dilution hypothesis (§4) predicts that reducing redundant non-text tokens should rebalance MDI. This section tests that prediction on the vision modality, where mature compression methods exist and where text dominance is both strong and well-characterized in our diagnostic results. We use existing methods as controlled tools to validate the attention dilution hypothesis and to assess the functional consequences of restoring attention balance (accuracy, latency).

5.1 Experimental Setup

We compare three token-reduction strategies under matched compression ratios (50%, 75%, 90%) on LLaVA-1.5 (7B and 13B) with MMMU-Pro:

- **Random dropping**: uniformly samples a subset of visual tokens, serving as a *mechanical baseline* that reduces token count without using any informativeness signal.
- **FastV** Chen et al. [2024]: prunes visual tokens at an intermediate LLM layer based on the attention they receive during the first few decoding steps.
- **FasterVLM** Zhang et al. [2024]: ranks visual tokens by the attention they receive from the `[CLS]` token in the final layer of the frozen vision encoder, and retains the top-ranked tokens before fusion with text embeddings.

This comparison separates two effects: matched compression ratios test whether reducing token count alone lowers MDI, while accuracy differences between random and informed selection indicate that *which* tokens are retained also matters beyond token count. Formal algorithmic details of the attention-guided selection procedure are provided in Appendix D.

Model	Reduction	Method	Early		Middle		Late		Acc	Time
			MDI	AEI	MDI	AEI	MDI	AEI		
LLaVA-7B	0 %	No reduction	1.91	1.55	8.07	2.92	15.63	3.36	33.23	0.42
	50 %	Random	1.38	1.30	4.14	1.84	8.98	2.28	31.30	0.15
		FastV	1.28	1.23	4.11	1.84	8.91	2.23	33.23	0.17
		FasterVLM	1.21	1.11	4.08	1.83	8.84	2.14	33.30 ↑	0.15
	75 %	Random	0.96	1.03	2.38	1.41	5.74	1.64	31.79	0.14
		FastV	0.95	1.01	2.41	1.44	5.78	1.66	33.17	0.16
		FasterVLM	0.92	0.97	2.29	1.32	5.77	1.55	33.58 ↑	0.14
	90 %	Random	0.82	0.87	1.38	1.15	3.49	1.29	30.29	0.12
		FastV	0.79	0.96	1.31	1.09	3.51	1.32	33.35	0.14
		FasterVLM	0.77	0.94	1.28	1.05	3.46	1.20	33.70 ↑	0.12
LLaVA-13B	0 %	No reduction	5.67	2.64	16.92	3.52	17.68	3.56	34.62	0.27
	50 %	Random	3.48	1.92	8.28	2.19	9.35	2.27	32.71	0.21
		FastV	3.25	1.80	8.14	2.15	9.24	2.20	34.52	0.22
		FasterVLM	3.09	1.68	8.02	2.11	9.13	2.15	34.86 ↑	0.21
	75 %	Random	2.34	1.39	4.72	1.59	6.20	1.66	32.15	0.19
		FastV	2.22	1.34	4.58	1.54	6.14	1.60	33.85	0.20
		FasterVLM	2.14	1.30	4.50	1.50	6.08	1.56	34.05 ↑	0.19
	90 %	Random	1.73	1.05	2.54	1.23	3.78	1.28	31.52	0.18
		FastV	1.65	1.12	2.43	1.18	3.72	1.23	33.90	0.19
		FasterVLM	1.60	1.10	2.37	1.15	3.68	1.20	34.28 ↑	0.18

Table 2: **Effect of token-reduction methods and ratios on MDI, AEI, accuracy, and latency** (LLaVA-1.5 on MMMU-Pro). 0% denotes the original setting. Under each non-zero ratio, we compare Random, FastV, and FasterVLM. ↑ indicates FasterVLM outperforms both Random and FastV in accuracy.

5.2 Results

Token compression rebalances MDI. As shown in Table 2, all three strategies reduce MDI monotonically with the compression ratio, supporting the view that token count is a major driver in this setting. At 90% compression on LLaVA-7B, the late-layer MDI drops from 15.63 (baseline) to 3.49 (random), 3.51 (FastV), and 3.46 (FasterVLM). The AEI exhibits a similar decreasing trend, indicating that the disproportionate attention advantage of text tokens diminishes as redundant visual tokens are removed.

Informed selection preserves accuracy better than random dropping. The gap between informed and random selection shows that token quality matters beyond token count. On LLaVA-7B, random dropping at 90% compression degrades accuracy to 30.29%, while FastV maintains 33.35% and FasterVLM achieves 33.70% (baseline: 33.23%).

Visual inspection of attention maps (Figure 5) illustrates that retained tokens correspond to semantically relevant regions.

Among the informed methods, the relative ranking varies with model size and compression ratio. FasterVLM achieves the highest accuracy across all tested ratios on both LLaVA-7B and LLaVA-13B, although the margins over FastV are small (e.g., 34.05 vs. 33.85 at 75% on LLaVA-13B) and should be interpreted cautiously. Compression also reduces latency substantially, from 0.42 s to 0.12 s on LLaVA-7B and from 0.27 s to 0.18 s on LLaVA-13B at 90% compression, while informed selection achieves comparable speed to random dropping at matched token budgets. Overall, in this vision setting, token compression substantially lowers MDI, and informed selection preserves accuracy better than random dropping while retaining comparable efficiency gains.

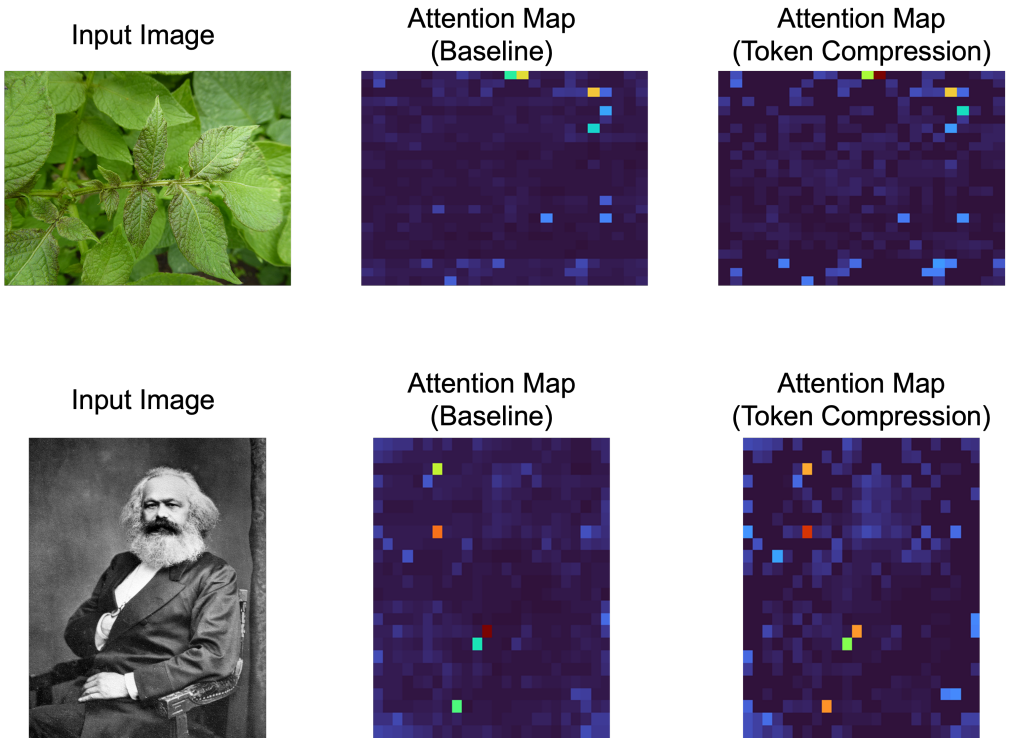

Figure 5: Attention maps under 50% token compression on two representative examples. From left to right: input image, baseline attention map, and attention map with 50% compression. In both cases, the baseline prediction is incorrect, while token compression yields the correct answer.

6 Discussion

Synthesizing the evidence. Our results across five modalities suggest that text dominance is a recurring pattern associated with the interaction between token count and information density. When non-text tokens are numerous and redundant (image, video, audio, time-series), per-token attention is diluted and text dominance emerges; graph tasks provide a boundary case where compact, information-dense non-text tokens can reverse the effect. In the image setting, the MDI bucket analysis (§3.3) further suggests a misalignment between attention allocation and task relevance: on LLaVA-1.5-7B with MMMU-Pro, the model tends to under-attend to image tokens more strongly on samples where removing images causes the largest accuracy drop. While this evidence is drawn from a single model–benchmark pair and should be interpreted with caution, it is consistent with the broader pattern observed across modalities.

Implications for MLLM design. The compression experiments in §5 show that, at least in the vision setting, removing up to 90% of visual tokens can lower MDI while maintaining accuracy. This suggests that the visual encoders tested here may produce more tokens than necessary for downstream reasoning, and that *token efficiency in non-text encoders may deserve attention alongside model scaling* as a design consideration. Whether this observation extends to other modalities and encoder families remains to be verified.

Scope and limitations. Several limitations should guide interpretation. MDI and AEI are diagnostic statistics of attention allocation, not causal measures of modality importance; we mitigate this by jointly reporting accuracy throughout the paper. Our token-replication experiments hold semantic content approximately fixed, but exact invariance cannot be guaranteed, as repeated tokens may affect layer normalization or positional encoding statistics. Cross-model comparisons (e.g., LLaVA vs. Qwen2.5-VL) are observational, since architecture, training data, and optimization co-vary.

7 Conclusion

We conducted a systematic, attention-level study of text dominance across five modalities in MLLMs, using two diagnostic metrics (MDI and AEI). Text dominance is pervasive across image, video, audio, and time-series settings, intensifies with non-text token count, and can reverse in graph tasks where non-text tokens are compact. On the vision modality, attention-guided token compression reduces MDI while preserving accuracy and lowering latency, supporting the connection between token redundancy and attention imbalance. Validating these findings on additional modalities, model families, and benchmarks is an important next step.

Acknowledgments

This work was supported in part by the NSFC Special Fund and Major Program under Grants 62550004 and 62595773; in part by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM101); in part by the CPSF (No. 2025M771546).

References

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

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

Weitong Cai, Jiabo Huang, Shaogang Gong, Hailin Jin, and Yang Liu. Mllm as video narrator: Mitigating modality imbalance in video moment retrieval. *Pattern Recognition*, 166:111670, 2025.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2024.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 3543–3556, 2019.

Hongrui Jia, Chaoya Jiang, Haiyang Xu, Wei Ye, Mengfan Dong, Ming Yan, Ji Zhang, Fei Huang, and Shikun Zhang. Symdpo: Boosting in-context learning of large multimodal models with symbol demonstration direct preference optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9361–9371, 2025.

Sicong Leng, Yun Xing, Zesen Cheng, Yang Zhou, Hang Zhang, Xin Li, Deli Zhao, Shijian Lu, Chunyan Miao, and Lidong Bing. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *arXiv*, 2024. URL `https://arxiv.org/abs/2410.12787`.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22195–22206, 2024.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409, 2024.

Chenxi Liu, Tianyi Xiong, Ruibo Chen, Yihan Wu, Junfeng Guo, Tianyi Zhou, and Heng Huang. Modality-balancing preference optimization of large multimodal models by adversarial negative mining. *arXiv preprint arXiv:2506.08022*, 2025.

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

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

Ying Liu, Ge Bai, Lu Chenji, Shilong Li, Zhang Zhang, Ruifang Liu, and Wenbin Guo. Eliminating the language bias for visual question answering with fine-grained causal intervention. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024b.

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

Xiaoyu Ma, Hao Chen, and Yongjian Deng. Improving multimodal learning balance and sufficiency through data remixing. *arXiv preprint arXiv:2506.11550*, 2025.

Jean Park, Kuk Jin Jang, Basam Alasaly, Sriharsha Mopidevi, Andrew Zolensky, Eric Eaton, Insup Lee, and Kevin Johnson. Assessing modality bias in video question answering benchmarks with multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19821–19829, 2025.

Yogesh Prabhu. *Unveiling Bias in Multimodal Models*. PhD thesis, 2025.

Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S Yu. A survey of multilingual large language models. *Patterns*, 6(1), 2025.

Jiabin Tang, Yuhao Yang, Wei Wei, Lei Shi, Lixin Su, Suqi Cheng, Dawei Yin, and Chao Huang. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 491–500, 2024.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. *arXiv preprint arXiv:2412.03104*, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *CoRR*, 2024.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

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*, pages arXiv–2412, 2024.

Xu Zheng, Chenfei Liao, Yuqian Fu, Kaiyu Lei, Yuanhuiyi Lyu, Lutao Jiang, Bin Ren, Jialei Chen, Jiawen Wang, Chengxin Li, et al. Mllms are deeply affected by modality bias. *arXiv preprint arXiv:2505.18657*, 2025.

A Experimental Environment

The system was equipped with an Intel(R) Xeon(R) Gold 6348 CPU operating at a base frequency of 2.60 GHz. For accelerating deep learning computations, an NVIDIA A800 Tensor Core GPU with 80 GB of dedicated memory was utilized. All models were evaluated using Python 3.12.3 and PyTorch 2.3.0 with CUDA 12.1 support.

B Data Preprocessing Parameters

To manage hardware limitations, data was preprocessed as follows. Images were dynamically resized to a total pixel count between $256 \times 28 \times 28$ and $1280 \times 28 \times 28$. For videos, processing was model-specific: VideoLLaMA3-7B frames were sampled at 1 fps up to a maximum of 90 frames and resized to 112×112; Qwen2.5-VL-7B frames were also sampled at 1 fps but resized to a maximum of 100×140 pixels.

C Modality Attention Calculation

To facilitate the calculation of the Modality Dominance Index (MDI) and the Attention Efficiency Index (AEI), we use the procedure detailed in Algorithm 1 to acquire the per-token attention scores for text and non-text inputs. This algorithm aggregates the attention weights from a specific output token to all input tokens, across designated layers and all their attention heads. After normalization and partitioning by modality, the process yields a text attention score S_{text} and a non-text attention score $S_{\text{non_text}}$ for that single output token. We first average the attention statistics over the generated tokens of each sample, then across samples. MDI is the ratio of the resulting mean per-token attention scores. AEI is computed from the mean attention masses and mean input token counts. For the Early, Middle, and Late statistics in Table 1, we use the first two, middle two, and last two model layers, respectively.

Algorithm 1 Modality Attention Calculation for an Output Token.

Require: Model M (with H heads); layer indices L; output token o; input text tokens $\mathcal{T}$; input non-text tokens $\mathcal{O}$.

Output: Text attention score S_{text}; non-text score $S_{\text{non_text}}$.

Let $A_i^{(k,h)}$ be the attention from o to input token i at layer $k \in L$, head h.

Compute mean attention: $\bar{a}_i = \frac{1}{|L| \cdot H} \sum_{k \in L} \sum_{h=1}^{H} A_i^{(k,h)}$.

Normalize: $a_{\text{norm}} = \bar{a} \,/\, \|\bar{a}\|_1$.

$S_{\text{text}} = \frac{1}{|\mathcal{T}|} \sum_{i \in \text{indices}(\mathcal{T})} (a_{\text{norm}})_i$.

$S_{\text{non_text}} = \frac{1}{|\mathcal{O}|} \sum_{i \in \text{indices}(\mathcal{O})} (a_{\text{norm}})_i$.

Return $S_{\text{text}},\ S_{\text{non_text}}$.

D Attention-Guided Visual Token Compression: Algorithmic Details

This appendix provides the formal procedure for the attention-guided token compression used in §5 (FasterVLM Zhang et al. [2024]).

Let the frozen vision encoder output a sequence of patch tokens $V = \{v_1, v_2, \ldots, v_N\}$ together with a class token [CLS]. For each patch token v_i, the saliency score is the normalized attention it receives from the class token in the final encoder layer:

$$s_i = \frac{1}{H} \sum_{h=1}^{H} \text{Attn}^{(h)}_{[\text{CLS}] \to i}, \qquad \tilde{s}_i = \frac{s_i}{\sum_{j=1}^{N} s_j}, \tag{5}$$

where H is the number of attention heads. Given a compression ratio $r \in [0, 1)$, we retain the top $M = \lfloor N(1-r) \rfloor$ tokens by saliency, forming the compressed set $V' = \{ v_i \mid \tilde{s}_i \geq \tilde{s}_{(M)} \}$, and discard the rest before fusion with textual embeddings. Algorithm 2 summarizes the procedure.

Algorithm 2 Attention-Guided Visual Token Compression Zhang et al. [2024].

Require: Vision encoder output $V = \{v_1, \ldots, v_N\}$; class token [CLS]; compression ratio r.

1: Compute per-head attention: $a_i^{(h)} = \mathrm{Attn}_{[\mathrm{CLS}] \to i}^{(h)}$.
2: Average across heads: $s_i = \frac{1}{H} \sum_{h=1}^{H} a_i^{(h)}$.
3: Normalize: $\tilde{s}_i = s_i \,/\, \sum_j s_j$.
4: Compute budget: $M = \lfloor N(1-r) \rfloor$.
5: Retain top-M tokens by $\tilde{s}_i$: $V' = \mathrm{TopK}(V, M)$.
6: **Return** compressed visual set V'.

E Model Attention and Token Statistics

This section presents a detailed analysis of attention distribution and token count metrics across various models and data modalities. To understand how these characteristics evolve through the model's depth, we segment our analysis by examining data from three distinct stages: the first two layers (Table 3), the middle two layers (Table 4), and the last two layers (Table 5). The tables summarize total attention and average attention per token for textual and non-textual inputs across the three layer groups.

Table 3: Attention and token statistics for the *first two layers*.

Modality	Model	Total Attn		Avg Attn/Tok		Avg Token Count		
		Text	Non-t	Text	Non-t	Text	Non-t	Out
Image	Qwen2.5-VL-7B	0.049	0.951	6.0e-4	2.7e-4	118.6	3584.0	208.0
	Kimi-VL-A3B	0.715	0.285	1.1e-2	1.3e-3	92.6	413.6	138.7
	LLaVA-1.5-7B	0.175	0.825	2.3e-3	1.4e-3	117.3	576.0	67.7
Video	Qwen2.5-VL-7B	0.092	0.908	2.7e-3	2.5e-4	34.8	3584.0	77.6
	VideoLLaMA3-7B	0.109	0.891	8.6e-3	4.5e-4	14.1	2288.4	15.8
Audio	Qwen2-Audio-7B	0.480	0.520	6.8e-3	6.7e-3	70.9	123.4	7.1
Time Series	ChatTS-14B	0.802	0.198	1.2e-2	7.9e-3	67.1	32.5	198.1
Graph	GraphGPT-7B	0.811	0.189	1.5e-3	1.1e-2	596.9	24.5	11.2

Table 4: Attention and token statistics for the *middle two layers*.

Modality	Model	Total Attn		Avg Attn/Tok		Avg Token Count		
		Text	Non-t	Text	Non-t	Text	Non-t	Out
Image	Qwen2.5-VL-7B	0.348	0.652	3.8e-3	1.8e-4	118.6	3584.0	208.0
	Kimi-VL-A3B	0.547	0.453	8.1e-3	2.1e-3	92.6	413.6	138.7
	LLaVA-1.5-7B	0.595	0.405	7.2e-3	7.0e-4	117.3	576.0	67.7
Video	Qwen2.5-VL-7B	0.423	0.577	1.2e-2	2.0e-4	36.7	3584.0	86.4
	VideoLLaMA3-7B	0.474	0.526	3.7e-2	2.6e-4	14.8	2283.0	15.8
Audio	Qwen2-Audio-7B	0.733	0.267	1.0e-2	3.2e-3	71.0	122.3	6.4
Time Series	ChatTS-14B	0.945	0.055	1.4e-2	3.2e-3	67.1	32.5	190.1
Graph	GraphGPT-7B	0.807	0.193	1.5e-3	1.0e-2	596.9	24.5	11.2

F Case Study: Text Dominance in Action

To provide qualitative intuition for the quantitative findings in the main text, Figure S1 contrasts two Biology questions from MMMU-Pro answered by Qwen2.5-VL-7B.

Table 5: Attention and token statistics for the *last two layers*.

Modality	Model	Total Attn		Avg Attn/Tok		Avg Token Count		
		Text	Non-t	Text	Non-t	Text	Non-t	Out
Image	Qwen2.5-VL-7B	0.456	0.544	5.0e-3	1.5e-4	118.6	3584.0	208.0
	Kimi-VL-A3B	0.474	0.526	7.0e-3	2.5e-3	92.6	413.6	138.7
	LLaVA-1.5-7B	0.716	0.284	8.6e-3	4.9e-4	117.3	576.0	67.7
Video	Qwen2.5-VL-7B	0.448	0.553	1.3e-2	1.5e-4	33.8	3584.0	83.8
	VideoLLaMA3-7B	0.491	0.509	3.7e-2	2.4e-4	15.2	2338.9	14.1
Audio	Qwen2-Audio-7B	0.507	0.493	7.1e-3	6.2e-3	71.0	80.0	20.0
Time Series	ChatTS-14B	0.924	0.076	1.4e-2	3.9e-3	67.1	32.5	191.0
Graph	GraphGPT-7B	0.860	0.140	1.6e-3	8.0e-3	596.9	24.5	11.2

In Case 1 (MDI = 2.3), the model allocates 54% of its attention to image tokens, successfully identifies the hair-like cilia in the micrograph, and answers correctly. In Case 2 (MDI = 7.1, a 3.1× increase), 72% of attention shifts to text tokens. Faced with a labeled neuron diagram, the model falls back on textual prior knowledge ("the myelin sheath is typically found around the axon") rather than reading the diagram labels, and incorrectly predicts B (axon) instead of the correct answer E.

Both questions belong to the same subject. The contrast illustrates the mechanism suggested by our quantitative analysis: when text dominance is high, the model substitutes textual priors for visual evidence, leading to errors on questions that require genuine image understanding.

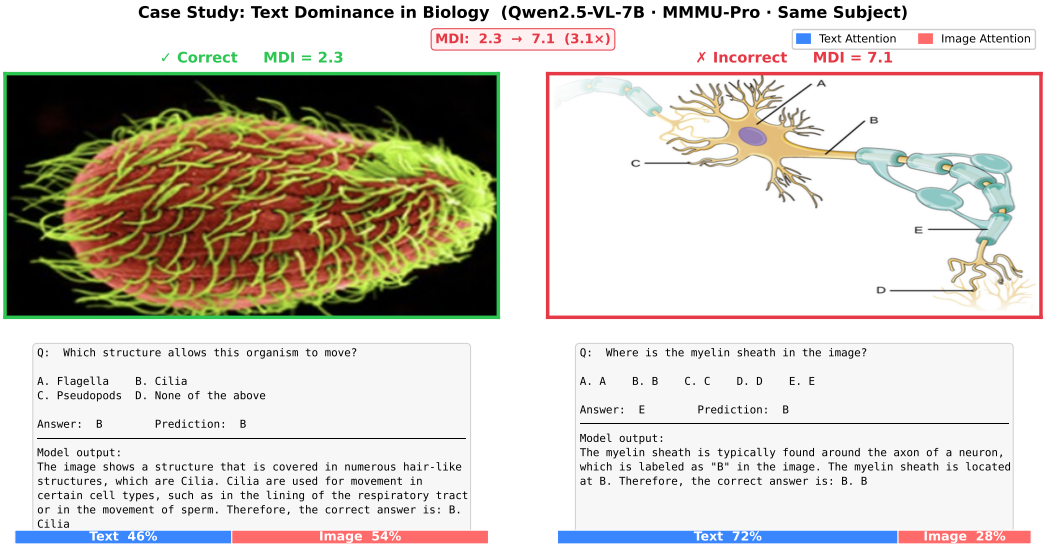

Figure S1: Case study of text dominance on two Biology questions (Qwen2.5-VL-7B, MMMU-Pro). **Left**: low MDI (2.3), the model attends to the image and answers correctly. **Right**: high MDI (7.1), the model over-relies on textual priors and misidentifies the labeled structure. The bottom bars show the text vs. image attention share.

G Additional Token-Compression Results

Within each experiment, all conditions are evaluated on the same items. Scores for random dropping and random selection average three draws.

G.1 Image Compression on MMBench-EN

We extend the image compression study to MMBench-EN [Liu et al., 2024c], where LLaVA-1.5-7B achieves a higher accuracy with full visual tokens. The evaluation uses 500 held-out items, with a

mean per-item chance accuracy of 27.38%. Informed selection and random dropping use matched visual-token budgets.

Table 6: Image compression on 500 MMBench-EN items.

Condition	Removed	Accuracy (%)
Full visual tokens	0%	73.4
Informed selection	50%	73.8
Informed selection	75%	73.6
Random dropping	75%	70.6
Informed selection	90%	72.6

Informed selection maintains accuracy close to the full-input result across the tested removal ratios. At 75% removal, its advantage over random dropping demonstrates the role of token quality at a fixed token count.

G.2 Video Compression on MVBench

We evaluate Qwen2.5-VL-7B on 2,000 randomly sampled MVBench items [Li et al., 2024]. At an intermediate LLM layer, informed selection retains the 50% of video tokens receiving the most attention. Random selection uses the same token budget and pruning layer. The retention ratio is fixed on a separate development split before evaluation.

Table 7: Video compression on 2,000 randomly sampled MVBench items.

Condition	Retained	Accuracy (%)	MDI
Full video	100%	61	523.83
Informed selection	50%	64	271.68
Random selection	50%	58	313.70

At the same token budget and pruning layer, informed selection outperforms random selection. Compared with the full-video baseline, informed selection achieves higher accuracy and a lower late-layer MDI, indicating a smaller ratio of average attention per text token to that per video token.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state claims that are directly supported by the experimental results in Sections 3–5. We distinguish between controlled evidence and observational evidence throughout.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated Limitations section discusses the non-causal nature of attention-based metrics, potential confounds in token replication, observational cross-model comparisons, and the scope of the compression experiments.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: This paper does not include theoretical results or proofs. The contributions are empirical.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: All model names, benchmark names, metric definitions, compression ratios, and algorithmic details (Algorithm 1) are provided. All benchmarks and models used are publicly available.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code for attention extraction, metric computation, and token compression will be released upon acceptance. All datasets and models used are publicly available.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Model names, benchmark splits, compression ratios, and layer groupings are specified in the main text. No training is performed; all experiments are inference-time evaluations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We report accuracy and attention metrics without error bars. Results for random dropping and random selection are averaged over three draws.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Hardware specifications (NVIDIA A800 80 GB GPU, Intel Xeon Gold 6348 CPU) and software versions are detailed in Appendix A. Inference latency is reported in Table 2. All experiments are inference-only and can be reproduced on a single GPU.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This work analyzes publicly available models and benchmarks. It does not involve human subjects, personal data, or dual-use concerns.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: This work is a diagnostic study of attention allocation in existing models. It does not introduce new models or capabilities that pose direct societal risks.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This paper does not release new models or datasets. The released code performs only diagnostic analysis on existing public models.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All models and benchmarks are cited with their original publications. All assets used are publicly available under permissive licenses.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new datasets or pretrained models. Evaluation code will be documented upon release.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: This paper does not involve human subjects research.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are the subject of study in this paper (we analyze their attention patterns), but they are not used as a non-standard component of our methodology.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2026/LLM`) for what should or should not be described.