

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

Huyu Wu¹,
 Meng Tang², Xinhan Zheng³,
 Haiyun Jiang^{4*}

¹Institute of Computing Technology, Chinese Academy of Sciences

²Department of Computer Science, Aberystwyth University

³Beijing University of Posts and Telecommunications

⁴Shanghai Jiao Tong University

huyu-wu@outlook.com, Met57@aber.ac.uk, chengfengke@bupt.edu.cn, haiyunjiangnlp@gmail.com

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated remarkable capabilities across a diverse range of multimodal tasks. However, these models suffer from a core problem known as text dominance: they depend heavily on text for their inference, while underutilizing other modalities. While prior work has acknowledged this phenomenon in vision-language tasks, often attributing it to data biases or model architectures. In this paper, we conduct the first systematic investigation of text dominance across diverse data modalities, including images, videos, audio, time-series, and graphs. To measure this imbalance, we propose two evaluation metrics: the Modality Dominance Index (MDI) and the Attention Efficiency Index (AEI). Our comprehensive analysis reveals that text dominance is both significant and pervasive across all tested modalities. Our in-depth analysis identifies three underlying causes: attention dilution from severe token redundancy in non-textual modalities, the influence of fusion architecture design, and task formulations that implicitly favor textual inputs. Furthermore, we propose a simple token compression method that effectively rebalances model attention. Applying this method to LLaVA-7B, for instance, drastically reduces its MDI from 10.23 to a well-balanced value of 0.86. Our analysis and methodological framework offer a foundation for the development of more equitable and comprehensive multimodal language models.

Introduction

Recent Multimodal Large Language Models (MLLMs) (Yin et al. 2024; Qin et al. 2025; Team et al. 2025; Bai et al. 2025) have achieved impressive success in both understanding and generation across diverse modalities, including images, videos, audio, and graph data. However, a critical weakness of these models is their modality imbalance (Cai et al. 2025; Zheng et al. 2025). A key limitation is MLLMs often disregard non-text inputs, generating outputs predominantly based on text context even when rich visual information is present (Jia et al. 2025).

This modality imbalance has been previously observed in tasks like Visual Question Answering (VQA). For instance,

some studies (Liu et al. 2024b) have shown that VQA models can often answer questions correctly even with the image absent, revealing a heavy reliance on linguistic priors. More recently, Leng et al. (Leng et al. 2024) proposed the Modality Importance Score (MIS) as a quantitative metric to evaluate modality imbalance in video question answering benchmarks. However, prior work has largely attributed this bias to data artifacts (Wang et al. 2024) or encoder design (Liu et al. 2024b; Luo et al. 2025), primarily within the image-text modality pair. The role of the internal attention mechanism, which is the very core of the Transformer architecture, in causing this imbalance, especially across a wider array of modalities, remains critically under-explored. This gap gives rise to a pivotal research question: *is text dominance a fundamental flaw of the Transformer architecture in MLLMs, extending beyond vision to modalities like audio, time-series, and graphs?*

To investigate this, we conduct the first systematic analysis of cross-modal attention in leading MLLMs across these five modalities. We introduce two novel metrics, the Modality Dominance Index (MDI) and the Attention Efficiency Index (AEI), to quantify this behavior. Our findings highlight a significant imbalance: in VideoLLaMA-7B, the MDI reaches 157, indicating that output tokens attend to text tokens 157 times more than to visual tokens on a per-token basis.

Through comprehensive analysis, we identify three principal factors contributing to text dominance. First, non-text modalities often contain excessive redundant tokens, which severely dilutes the model’s attention. Second, complex multimodal fusion architectures tend to amplify this imbalance, whereas more straightforward fusion designs facilitate a more balanced allocation of attention. Third, many multimodal tasks formulations naturally privilege text inputs, naturally guiding the model to focus more heavily on the text modality.

Motivated by our finding on attention dilution, we propose a simple yet effective solution: token compression. By strategically reducing redundant tokens within non-text modalities, this approach substantially rebalances cross-modal attention distributions. This method enhances the in-

*Corresponding author

formation density per token and effectively mitigates text dominance.

On this basis, our contributions are as follows:

- We provide the first evidence that text dominance is a fundamental and pervasive bias in Transformer-based MLLMs, extending across a wide spectrum of modalities.
- We conduct a comprehensive analysis of the underlying causes, including token redundancy in non-text modalities, the influence of fusion architecture design, and task formulations that implicitly favor textual inputs.
- We present and validate token compression, a straightforward and effective approach to mitigate text dominance.

The main content of this paper is presented in the following sections. Section 3 details the evaluation framework and formalizes our core metrics. Section 4 offers a comprehensive analysis of text dominance, considering different model architectures and the impact of task design across multiple modalities. Section 5 describes the token compression approach and examines its effectiveness in addressing modality imbalance.

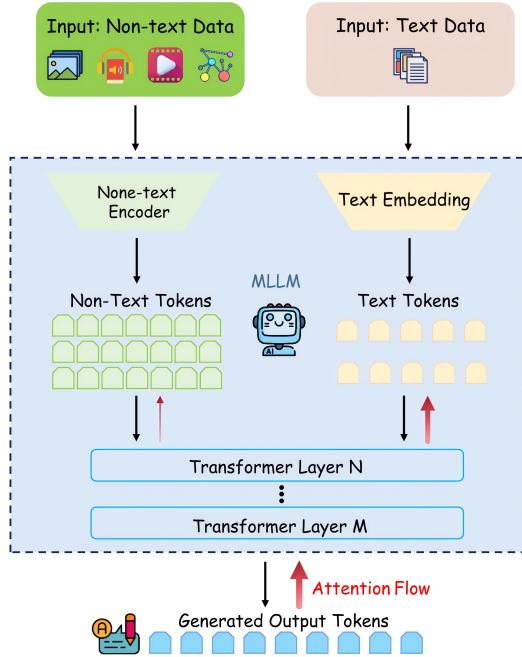


Figure 1: Each modality is tokenized and jointly processed by the MLLM. The red arrows illustrate the attention mechanism among non-text, text, and generated output tokens. The thinner arrows associated with non-text modalities reflect their larger token count and, consequently, the lower per-token attention weights.

Related Work

The Expanding Frontier of Multimodal Large Language Models

The remarkable success of Large Language Models (LLMs) (Yin et al. 2024; Kumar 2024) has catalyzed a paradigm shift towards Multimodal Large Language Models (MLLMs) (Yin et al. 2024; Qin et al. 2025), which integrate diverse data modalities. The canonical MLLM architecture comprises a pretrained modality-specific encoder, a powerful LLM serving as the cognitive core, and a carefully designed interface to align representations across modalities (Liang et al. 2024).

Building on this foundation, researchers have rapidly extended the capabilities of MLLMs beyond images, to capture spatio-temporal dynamics in videos, models like Video-LLaMA (Zhang et al. 2025) and its successors incorporate specialized components to explicitly model temporal dependencies and fuse audiovisual signals. For audio, models like Qwen-Audio (Chu et al. 2024) adopt tokenization via Vector Quantization (VQ) to convert continuous waveforms into discrete sequences compatible with LLMs. The exploration has further ventured into sequential and structured data. For instance, models like Chat-TS (Xie et al. 2024) have been developed to handle complex time-series data by encoding temporal patterns into the LLM’s latent space. In the realm of graph-structured data, GraphGPT (Tang et al. 2024) demonstrates the potential of LLMs to comprehend and reason over relational information by translating graph structures into a format that LLMs can process.

Modality Imbalance in Multimodal Large Language Models

The phenomenon of modality imbalance (Prabhu 2025) refers to the model’s tendency to over-rely on text while underutilizing or entirely ignoring information from another modality, such as vision.

The roots of modality imbalance can be traced to both data and model architecture. First, inherent data bias is a primary contributor, as the higher information density of text compared to complex image data creates an exploitable shortcut for the model (Park et al. 2025). Second, the architectural design of MLLMs systematically exacerbates this imbalance. Most MLLMs exhibit Asymmetric Modal Backbone Capabilities, coupling an immensely powerful LLM pretrained on trillions of text tokens with a vision encoder trained on a comparatively smaller scale of data (Li et al. 2023; Liu et al. 2023).

Strategies for Mitigating Modality Imbalance

To address modality imbalance, the research community has proposed mitigation strategies from multiple perspectives. One line of work focuses on redesigning the training process at the data level to proactively prevent the imbalance. The Data Remixing framework (Ma, Chen, and Deng 2025) introduces a two-stage training strategy. It performs sample-level decoupling by masking the stronger modality, which forces the model to rely on the weaker one and counteracts modality inertia. A recent approach, the MBPO framework

(Liu et al. 2025), directly targets the model’s over-reliance on text. It employs Direct Preference Optimization (DPO) on adversarially generated “hard negatives” to compel the model to favor visual evidence over language-driven hallucinations.

Text Dominance in Multimodal Large Language Models

Overview

The rapid development of Multimodal Large Language Models (MLLMs) has demonstrated their remarkable abilities in multimodal understanding and inferencing. Although these models are theoretically capable of integrating information from modalities such as text, images, video, audio, time-series data, and graphs, a persistent challenge has emerged: During generation, MLLMs commonly give greater weight to text over non-textual modalities. This phenomenon, referred to as text modality dominance, is marked by the model allocating substantially more attentional resources to textual content compared to other modalities.

While this phenomenon is primarily documented within the vision-language domain, we propose this dominance also exists in video, audio, time-series, and graph modalities. However, systematic cross-modal empirical validation is lacking.

To address this issue, we propose a series of token-level analyses leveraging the cross-attention mechanism inherent in generative MLLMs. Specifically, we leverage the cross-attention mechanisms employed by MLLMs during the generation process, quantitatively analyzing the attention distribution between output tokens and input tokens across different modalities.

This enables a direct statistical measurement: we compare the proportion of attention allocated to textual inputs against that allocated to non-textual inputs. The resulting metric provides a quantitative and interpretable assessment of text modality dominance.

Datasets and Baselines

To construct a comprehensive and robust evaluation framework, we selected representative datasets and state-of-the-art models for five key modalities, including image, video, audio, time-series, and graph. For the image modality, we employed the MMMU-Pro benchmark (Yue et al. 2024), which excludes questions answerable by text alone, assessing visual-text fusion. We evaluated three state-of-the-art vision-language models on this task: Qwen2.5-VL-7B (Bai et al. 2025), LLaVA-1.5-7B (Liu et al. 2024a), and Kimi-VL-A3B-Instruct (Team et al. 2025), each representing different multimodal architectures.

For video analysis, the MMBench-Video benchmark (Fang et al. 2024) assesses temporal reasoning on YouTube long-form content with open-ended questions. Our evaluation on this benchmark included two distinct models: Qwen2.5-VL-7B (Bai et al. 2025), a general-purpose model adapted from an image-text foundation, and VideоЛLaMA3-7B (Zhang et al. 2025), a specialist model explicitly optimized for video-centric tasks.

For audio, the IEMOCAP dataset (Busso et al. 2008), with multi-turn annotated conversations, was used to test Qwen2-Audio-7B-Instruct (Chu et al. 2024), a language model with integrated speech encoding.

In time-series, we evaluated ChatTS-14B (Xie et al. 2024), designed for multivariate temporal reasoning, on synthetic tasks, focusing on attention balance between text and time-series data.

For graph data, we employ GraphGPT-7B (Tang et al. 2024) and its corresponding benchmark, GraphGPT-eval-instruction. This framework aligns a large language model with graph knowledge through a two-stage, instruction fine-tuning paradigm. We conduct inference tests using its instruction set to measure the model’s attention allocation across graph information.

Evaluation Metrics

To characterise how a 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.

Modality Dominance Index. The MDI quantifies the relative reliance of a multimodal model on textual versus non-textual inputs during autoregressive generation. For an input sequence comprising a set of textual tokens \mathcal{T} and a set of non-textual tokens \mathcal{O} , we first compute the total attention scores directed towards each modality. Let A_T and A_O be the attention scores aggregated over the generation of N output tokens for all tokens in \mathcal{T} and \mathcal{O} respectively, normalized such that $A_T + A_O = 1$. The MDI is then formulated as the ratio of the average per-token attention between the two modalities:

$$\text{MDI} = \left(\frac{A_T}{|\mathcal{T}|} \right) \cdot \left(\frac{A_O}{|\mathcal{O}|} \right)^{-1} \quad (1)$$

Thus, MDI values above 1 signify text dominance; values below 1 indicate non-text dominance; and values close to 1 correspond to a balanced influence from both.

Attention Efficiency Index. To complement the MDI, we introduce the Attention Efficiency Index (AEI), which considers the computational resources consumed by each modality. While most existing metrics focus on absolute attentional dominance, they often overlook costs such as token allocation across modalities. The AEI measures the efficiency of a modality in converting its token representation into attention, providing a normalized assessment of resource usage in multimodal generation.

Let A_T be the total attention score for text tokens and A_O for non-text tokens. The proportion of attention captured by the text modality, P_T , is:

$$P_T = \frac{A_T}{A_T + A_O} \quad (2)$$

Given $|\mathcal{T}|$ text tokens and $|\mathcal{O}|$ non-text tokens, the proportional size of the text modality in the input, Q_T , is:

$$Q_T = \frac{|\mathcal{T}|}{|\mathcal{T}| + |\mathcal{O}|} \quad (3)$$

The AEI for the text modality is then defined as the ratio of its attention share to its token share:

$$\text{AEI}_T = \frac{P_T}{Q_T} = \frac{A_T/(A_T + A_O)}{|\mathcal{T}|/(|\mathcal{T}| + |\mathcal{O}|)} \quad (4)$$

An AEI value greater than 1 indicates high efficiency, signifying that the modality achieves disproportionate attentional prominence relative to its token allocation. By distinguishing absolute dominance from resource efficiency, the AEI quantifies how effectively a modality leverages its token representation to influence the model’s attentional mechanisms.

Together, MDI and AEI allow us to disentangle *dominance* from *efficiency*: MDI assesses which modality ultimately governs the generation process, while AEI evaluates how productively a modality uses its limited token budget to capture the model’s focus.

Experimental Results

To quantify attention allocation in MLLMs, we analyze the Modality Dominance Index (MDI) and Attention Efficiency Index (AEI) across different model layers. As detailed in Table 1, our measurements reveal a clear and consistent pattern across various models, modalities, and benchmarks: regardless of layer depth, *textual dominance is evident*, though its degree varies, often intensifying in deeper layers for certain tasks while remaining stable or moderate in others.

This hierarchical trend towards text dominance is particularly pronounced in the mainstream modalities of image and video. For the Qwen2.5-VL-7B model on the image modality, the MDI rises from 2.26 in early layers to 33.10 in late layers. This signifies that in the later stages of processing, the average attention allocated to each text token is over 33 times greater than that given to an image token. Meanwhile, the AEI drops from 14.24 to 1.42, illustrating the shift in attention allocation. In video tasks, VideоЛLaMA3-7B reaches a late-layer MDI of 157.53 on the MMBench-Video benchmark, indicating that text tokens attract over two orders of magnitude more attention than video frame tokens.

We further investigated the effect of non-textual information volume on attention allocation through controlled experiments. In audio and time-series tasks, we kept the text input constant while replicating the non-textual token sequence fivefold and tenfold. The data shows that this change in input scale systematically exacerbates text dominance. For Qwen2-Audio-7B-Instruct, the late-layer MDI increases from an initial 1.16 to 6.73 and 8.70 as the replication factor grows. Similarly, the late-layer MDI for ChatTS-14B climbs from 3.52 to 9.28 and 16.25. These results indicate that as the proportion of non-textual tokens in the input increases, the model’s relative focus on text grows disproportionately.

Conversely, tasks involving graph modalities present an initial exception. For GraphGPT-7B under standard conditions, late-layer MDI is 0.20, indicating preference for the non-textual graph modality. Yet, with 10-fold replication of non-textual tokens, MDI rises to 1.35, exceeding the equilibrium threshold of 1.0 and denoting a shift to textual modality dominance. This suggests that such dominance can arise

even in initially non-text-favoring models under altered input ratios.

In summary, our layer-wise evaluation of MDI and AEI confirms the prevalence of text modality dominance in MLLMs. This dominance often strengthens in deeper layers for many tasks, though the pattern varies by modality and input conditions. It appears across modalities such as image and video, increases with higher non-textual token proportions, and may arise even in tasks that initially favor non-textual modalities, as observed in graph-based examples under token replication. These findings provide a foundation for exploring causal mechanisms in the next section.

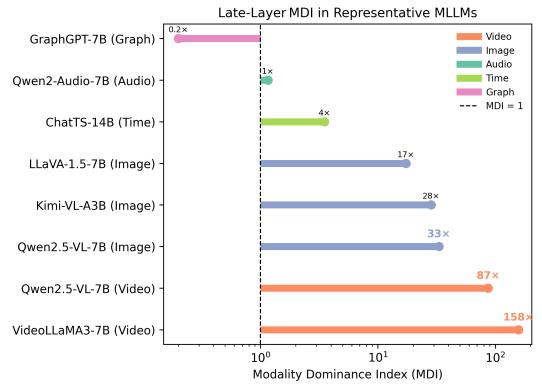


Figure 2: Text dominance phenomenon across MLLMs of different modalities. The dashed line in the figure marks MDI = 1; points situated to its right demonstrate a text-dominant pattern. The graph modality falls to the left of this threshold, and we provide a dedicated explanation for this observation in the analysis section.

Causes of Text Dominance

Multimodal Large Language Models (MLLMs) have demonstrated remarkable performance across tasks involving images, video, audio, and time-series data. However, a recurring phenomenon known as text dominance has emerged: during inference, MLLMs tend to overemphasize textual tokens while underutilizing non-textual modalities. Figure 2 shows that this pattern appears across various model architectures. While previous studies have attributed this to inherent modality priors or alignment biases introduced during pretraining, we propose a different explanation. Our findings suggest that text dominance is not a reflection of static modality preferences but rather a dynamic consequence of token-level imbalance across modalities, which leads to a phenomenon we refer to as attention dilution.

Token Redundancy Drives Attention Dilution

To systematically investigate the underlying causes of text dominance in MLLMs, we present a thorough analysis of the rising number of tokens and resulting attention dilution during the encoding phase in widely adopted multimodal architectures. Our study reveals that non-text modalities have redundant tokens, reducing their effectiveness in cross-modal

Model	Modality	Dataset	Early		Middle		Late	
			MDI	AEI	MDI	AEI	MDI	AEI
Qwen2.5-VL-7B	Image	MMMU_Pro	2.26	14.24	21.12	10.86	33.10	1.42
			3.84	2.82	54.96	21.88	26.03	13.95
			9.33	6.15	92.21	60.43	24.46	14.60
			1.58	1.04	10.23	3.51	17.37	4.23
			2.27	3.91	3.78	2.99	28.39	2.59
Kimi-VL-A3B-Instruct	Video	MMBench-Video	10.72	9.60	74.13	41.78	86.95	47.84
			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
			2.65	2.56	8.09	5.17	6.73	4.31
			2.80	2.50	10.10	5.46	8.70	5.09
ChatTS-14B	Time-series	TimeSeries-Reasoning $\times 1$	1.52	1.19	4.37	1.40	3.52	1.37
			2.08	1.95	10.72	3.15	9.28	3.03
			2.36	2.67	20.70	5.37	16.25	5.13
GraphGPT-7B	Graph	GraphGPT-Eval-Instruction $\times 1$	0.14	0.84	0.14	0.84	0.20	0.90
			0.20	0.69	0.35	0.83	0.69	0.98
			0.31	0.71	0.68	0.97	1.35	1.14

Table 1: Comparative analysis of the Modality Dominance Index (MDI) and Attention Efficiency Index (AEI) across diverse models, modalities, and benchmarks. The notation “ $\times n$ ” represents the replication factor applied to tokens from non-textual modalities. “Early,” “Middle,” and “Late” denote aggregated statistics from the first two, middle two, and last two model layers, respectively.

attention computation.

Concretely, video inputs are processed as extended sequences of frames, while audio and time-series data are commonly partitioned into numerous patches or temporal segments. Such preprocessing steps inevitably lead to a significant rise in the number of tokens for non-text modalities. As a result, these tokens tend to be highly redundant and exhibit relatively low semantic density. In contrast, text tokens are semantically compact and contain concentrated semantic information.

Due to this imbalance, the attention mechanism tends to prioritize textual tokens, causing pronounced attention dilution in non-text modalities. For example, on the MMBench-Video benchmark, Video-LLaMA3-7B demonstrates a Modality Dominance Index (MDI) of 157.53 in the model’s late layers, indicating that each text token receives on average over 157 times the attention weight assigned to an individual video frame token during generation. Correspondingly, the Attention Efficiency Index (AEI) achieves a value of 76.26, highlighting that text tokens, while comprising only a small portion of the total input, receive a disproportionately large share of the model’s attention. This reveals an imbalance in attention allocation within MLLMs: even when non-text inputs make up the majority of tokens, the models still primarily rely on textual information during inference. As a result, video frames and other non-text tokens are effectively marginalized within the competitive attention mechanism, potentially limiting the model’s ability to fully exploit multimodal information.

Fusion Architecture Impact on Text Dominance

Beyond the token structure of input modalities, architectural design critically shapes how attention is distributed and which modality dominates during inference. As illustrated in Figure 3, we conduct a comparative analysis of the MDI

and AEI between two representative vision-language multi-modal models.

LLaVA-1.5 7b uses a shallow bridging architecture with a frozen visual encoder and linear projection module, where the MDI for vision tasks rises from 1.58 in the early layers to 17.37 in the later layers. In contrast, Qwen2.5-VL employs a more integrated fusion mechanism featuring a Vision Transformer encoder combined with an MLP-based vision-language merger module, leading to a markedly higher modality dominance index at corresponding stages, reaching as high as 33.1. This suggests that deeper fusion mechanisms can amplify the dominance of the textual modality to a certain extent.

However, from the perspective of AEI, LLaVA-1.5 maintains a relatively high and increasing AEI, rising from 1.03 to 4.23, whereas Qwen2.5-VL exhibits a continuous decline in AEI, dropping from 14.24 in the early layers to 1.42 in the late layers. This phenomenon highlights a noteworthy trade-off: while complex architectures may enhance textual control, they potentially compromise overall attention utilization efficiency. Conversely, simpler architectures, under constrained resource allocation, encourage more efficient use of textual inputs, thereby achieving a novel balance between control and attention efficiency. These insights provide valuable guidance for future model design, emphasizing the need to balance enhanced modality representation capacity with optimized attention resource allocation.

Text Modality Leads Attention in Task Design

Furthermore, beyond architectural and representational factors, task design itself can profoundly influence attention allocation across modalities. In certain tasks, the shift in attention towards the textual modality arises not solely from differences in input representation, but more fundamentally from structural dependencies on textual prompts embedded within the task formulation. For instance, in time-series

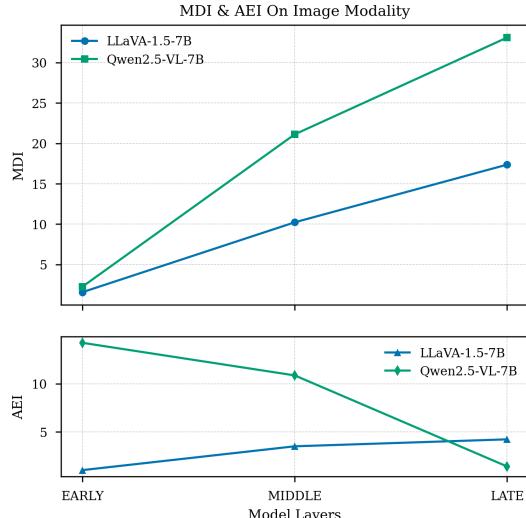


Figure 3: MDI and AEI comparison between LLaVA-1.5-7B and Qwen2.5-VL-7B on the image modality across stages.

tasks, key normalization factors and task-specific metadata are often encoded in natural language instructions, establishing the textual modality’s logical dominance from the input stage. Similarly, in audio-related tasks such as emotion recognition or keyword alignment, the task objective is typically guided by textual prompts, placing the textual modality at the semantic and inferential core.

To further validate this phenomenon, we analyze two representative models ChatTS-14B and Qwen2-Audio-7B under varying levels of non-text token replication ($\times 1$, $\times 5$, $\times 10$), examining their modality-specific attention distribution, as shown in Figure 4. Remarkably, even without expanding non-text tokens ($\times 1$ configuration), the textual modality consistently exhibits a clear advantage in attention allocation: ChatTS-14B achieves an Attention Efficiency Index (AEI) of 1.37 at the late layers, while Qwen2-Audio-7B reaches an AEI of 1.08 in the same stage.

As the quantity of non-text tokens increases, the dominance of the textual modality not only persists but becomes increasingly pronounced. Specifically, for ChatTS-14B, the MDI rises markedly from 4.37 at the middle layers under the single replication setting to 10.72 at the middle layers with fivefold replication, and further surges to 20.70 at the late layers under tenfold replication. Correspondingly, its AEI increases from 1.37 in the late layers of the single replication configuration to 3.03 and 5.13 at the late layers for fivefold and tenfold replications, respectively. A similar pattern is observed with Qwen2-Audio-7B, where the MDI ascends from 3.24 at the middle layers with single replication to 10.10 at the middle layers under tenfold replication. Simultaneously, its AEI escalates from 1.08 at the late layers in the single replication setting to 5.09 at the middle layers with tenfold replication.

These findings, supported by the observed trends, provide strong evidence that in tasks with a high reliance on textual prompts, models consistently prioritize attention allocation

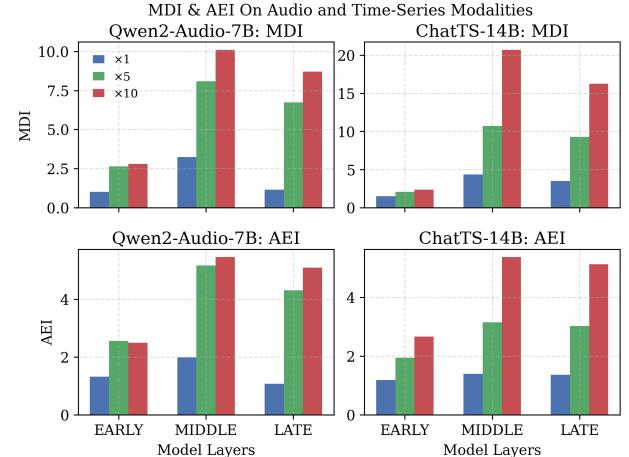


Figure 4: MDI and AEI of Audio and Time-Series Models with Token Scaling.

toward text tokens, even when non-text modalities are more numerous. These results indicate that textual prompts play a key role in directing attention and inference in multimodal language models.

Modality Dominance Shift in Graph-Based Tasks

Contrary to the prevailing trend of textual modality dominance observed in multimodal models, the performance of GraphGPT on graph-related tasks presents a notable exception. When the graph input is relatively small and the number of graph tokens is substantially lower than that of the accompanying textual prompt, the model’s MDI is initially measured at 0.20. This low value indicates that, on average, graph tokens attract more attention than textual tokens in this configuration. At the same time, the AEI for the textual modality remains at 0.90, suggesting that the textual input is neither dominant nor particularly effective in garnering attention resources under these conditions.

Under such conditions, the model naturally allocates more attention to the information-dense graph tokens, reflecting an inherent preference for inputs with higher semantic compactness, irrespective of modality. To further probe this behavior, we systematically increased the number of graph tokens by replication—scaling them by 5 \times and 10 \times without altering their semantic content. As a result, the MDI increased from 0.20 to 1.35, and the textual AEI rose from 0.90 to 1.14. These shifts indicate a gradual transition in modality dominance from graph to text, accompanied by a corresponding increase in the attention efficiency of textual tokens—from below-baseline to above-baseline.

This controlled modulation provides compelling empirical support for our central hypothesis: modality dominance is not a fixed characteristic encoded by pretraining, but a dynamic response driven by the structure and statistics of the input. The model’s allocation of attention across modalities is primarily governed by token count and information density, rather than by any static or modality-specific prior. In this light, observed modality preferences emerge

Method	Reduction	Early		Middle		Late	
		MDI	AEI	MDI	AEI	MDI	AEI
LLaVA-1.5-7B	0 %	1.58	1.04	10.23	3.51	17.37	4.23
FasterVLM	75 %	0.57	0.71	1.81	1.33	3.39	1.64
	90 %	0.57	0.80	1.10	1.03	1.84	1.17
	95 %	0.48	0.82	0.86	0.97	3.39	1.64

Table 2: **Effect of token-reduction ratio on Modality Dominance Index (MDI) and Attention Efficiency Index (AEI).** Statistics are reported for the first two layers (Early), the middle two layers (Middle), and the final two layers (Late).

as input-induced and context-sensitive outcomes, rather than immutable architectural biases.

Token Compression for Text Dominance

Building on the finding of attention dilution phenomenon, we propose optimization strategies for current architectures to rebalance modality integration. Our results show that when multimodal information is combined with textual input, text dominance tends to intensify. For example, in LLaVA-1.5-7B, MDI rises to 17.37 in the later layers, indicating that each text token receives over 17 times the attention of a single visual token on average. This highlights an imbalance in token utilization: while text inputs remain semantically dense despite a relatively small number of tokens, a single image is usually represented by hundreds of visual tokens, many of which are redundant or carry low informational value.

To address text modality dominance, we build on recent work by utilizing the [CLS] token attention mechanism (Zhang et al. 2024) derived from a frozen visual encoder as a more reliable indicator for visual token pruning. The [CLS] token is designed to capture the global semantics of the image via self-attention and provides stable visual token saliency assessments consistent across network layers. Formally, given N visual tokens $V = \{v_1, \dots, v_N\}$ encoded by a visual transformer, we compute the importance score s_i for each token v_i as

$$s_i = \text{Attn}([\text{CLS}], v_i). \quad (5)$$

Then, applying a token reduction rate r , only the top

$$M = N(1 - r) \quad (6)$$

tokens with the highest scores are retained, forming a compressed sequence

$$V' = \{v'_1, \dots, v'_M\}. \quad (7)$$

This [CLS]-guided compression strategy directly mitigates attention dilution by reducing the cardinality of non-textual inputs $|\mathcal{O}|$, thereby rebalancing the allocation of attention across modalities. The pruning threshold τ is adaptively determined according to a given computational budget R as follows:

$$\tau = \min \{ \tau \mid |\{a \in a_{[\text{CLS}]} \mid a \geq \tau\}| \leq N \times (1 - R) \} \quad (8)$$

where $a_{[\text{CLS}]}$ represents the attention scores from the [CLS] token.

We conducted experiments on the LLaVA-1.5-7B model using the MMMU Pro benchmark, evaluating both the MDI and AEI at early, middle, and late network layers under different compression rates: 0%, 75%, 90%, and 95%. The results are reported under the method name FasterVLM, which applies [CLS]-guided token pruning to reduce redundant visual tokens before fusion. As shown in Table 2, increasing the compression rate from 0% to 90% leads to a substantial reduction in the late-layer MDI, dropping from 17.37 to 1.84. This effectively alleviates text modality dominance and brings the attention distribution closer to balance, as MDI approaches one. This result demonstrates that compressing non-text input tokens allows the model to make better use of visual information.

Further analysis shows that as MDI decreases, the AEI for the text modality also declines from 4.23 to 1.17. This indicates a shift from strong reliance on text input towards a more balanced integration of different modalities. These results support our main hypothesis that text dominance can be influenced by adjusting the input structure. By reducing the number of non-text tokens in an appropriate way, the model’s focus can be redistributed to enable more balanced multimodal inferencing.

Additionally, our work extends the scope of prior research (Zhang et al. 2024), demonstrating that token compression techniques not only enhance computational efficiency but also play a significant role in alleviating text modality dominance. Together, these results contribute practical strategies for balancing modality integration and offer a clearer characterization of attention distribution mechanisms within MLLMs.

Conclusion

In this work, we systematically examined the phenomenon of text dominance in Multimodal Large Language Models. We introduced two metrics, the Modality Dominance Index (MDI) and the Attention Efficiency Index (AEI), to measure and analyze how attention is allocated among different input modalities. Experiments on images, video, audio, time-series, and graph data demonstrate that text modality dominance is common in current models. We also found that compressing non-text tokens mitigates this imbalance and facilitates more equitable multimodal integration. These results provide valuable tools and guidance for building more

efficient and balanced multimodal models.

Future work will explore additional strategies such as architectural redesign to foster more integrated modality fusion and task reformulation to reduce over-reliance on textual prompts. These approaches will be systematically investigated to evaluate their effectiveness and potential synergy with token compression, aiming to advance the development of robust and balanced multimodal foundation models. Through these methods, we aim to mitigate text dominance and maximize the utilization of multimodal information.

References

- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4): 335–359.
- Cai, W.; Huang, J.; Gong, S.; Jin, H.; and Liu, Y. 2025. MLLM as video narrator: Mitigating modality imbalance in video moment retrieval. *Pattern Recognition*, 166: 111670.
- Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; et al. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.
- Fang, X.; Mao, K.; Duan, H.; Zhao, X.; Li, Y.; Lin, D.; and Chen, K. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37: 89098–89124.
- Jia, H.; Jiang, C.; Xu, H.; Ye, W.; Dong, M.; Yan, M.; Zhang, J.; Huang, F.; and Zhang, S. 2025. 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*, 9361–9371.
- Kumar, P. 2024. Large language models (LLMs): survey, technical frameworks, and future challenges. *Artificial Intelligence Review*, 57(10): 260.
- Leng, S.; Xing, Y.; Cheng, Z.; Zhou, Y.; Zhang, H.; Li, X.; Zhao, D.; Lu, S.; Miao, C.; and Bing, L. 2024. The Curse of Multi-Modalities: Evaluating Hallucinations of Large Multimodal Models across Language, Visual, and Audio. *arXiv*.
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.
- Liang, Z.; Xu, Y.; Hong, Y.; Shang, P.; Wang, Q.; Fu, Q.; and Liu, K. 2024. A survey of multmodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, 405–409.
- Liu, C.; Xiong, T.; Chen, R.; Wu, Y.; Guo, J.; Zhou, T.; and Huang, H. 2025. Modality-Balancing Preference Optimization of Large Multimodal Models by Adversarial Negative Mining. *arXiv preprint arXiv:2506.08022*.
- Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26296–26306.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.
- Liu, Y.; Bai, G.; Chenji, L.; Li, S.; Zhang, Z.; Liu, R.; and Guo, W. 2024b. Eliminating the Language Bias for Visual Question Answering with fine-grained Causal Intervention. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6. IEEE.
- Luo, G.; Yang, X.; Dou, W.; Wang, Z.; Liu, J.; Dai, J.; Qiao, Y.; and Zhu, X. 2025. Mono-internvl: Pushing the boundaries of monolithic multimodal large language models with endogenous visual pre-training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 24960–24971.
- Ma, X.; Chen, H.; and Deng, Y. 2025. Improving Multimodal Learning Balance and Sufficiency through Data Remixing. *arXiv preprint arXiv:2506.11550*.
- Park, J.; Jang, K. J.; Alasaly, B.; Mopidevi, S.; Zolensky, A.; Eaton, E.; Lee, I.; and Johnson, K. 2025. Assessing modality bias in video question answering benchmarks with multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19821–19829.
- Prabhu, Y. 2025. *Unveiling Bias in Multimodal Models*. Ph.D. thesis.
- Qin, L.; Chen, Q.; Zhou, Y.; Chen, Z.; Li, Y.; Liao, L.; Li, M.; Che, W.; and Yu, P. S. 2025. A survey of multilingual large language models. *Patterns*, 6(1).
- Tang, J.; Yang, Y.; Wei, W.; Shi, L.; Su, L.; Cheng, S.; Yin, D.; and Huang, C. 2024. Graphgpt: Graph instruction tuning for large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 491–500.
- Team, K.; Du, A.; Yin, B.; Xing, B.; Qu, B.; Wang, B.; Chen, C.; Zhang, C.; Du, C.; Wei, C.; et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.
- Wang, W.; Chen, Z.; Wang, W.; Cao, Y.; Liu, Y.; Gao, Z.; Zhu, J.; Zhu, X.; Lu, L.; Qiao, Y.; et al. 2024. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *arXiv preprint arXiv:2411.10442*.
- Xie, Z.; Li, Z.; He, X.; Xu, L.; Wen, X.; Zhang, T.; Chen, J.; Shi, R.; and Pei, D. 2024. Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. *arXiv preprint arXiv:2412.03104*.
- Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2024. A survey on multimodal large language models. *National Science Review*, 11(12): nwae403.
- Yue, X.; Zheng, T.; Ni, Y.; Wang, Y.; Zhang, K.; Tong, S.; Sun, Y.; Yu, B.; Zhang, G.; Sun, H.; et al. 2024. MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark. *CoRR*.

Zhang, B.; Li, K.; Cheng, Z.; Hu, Z.; Yuan, Y.; Chen, G.; Leng, S.; Jiang, Y.; Zhang, H.; Li, X.; et al. 2025. Videolama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*.

Zhang, Q.; Cheng, A.; Lu, M.; Zhuo, Z.; Wang, M.; Cao, J.; Guo, S.; She, Q.; and Zhang, S. 2024. [CLS] Attention is All You Need for Training-Free Visual Token Pruning: Make VLM Inference Faster. *arXiv e-prints*, arXiv–2412.

Zheng, X.; Liao, C.; Fu, Y.; Lei, K.; Lyu, Y.; Jiang, L.; Ren, B.; Chen, J.; Wang, J.; Li, C.; et al. 2025. MLLMs are Deeply Affected by Modality Bias. *arXiv preprint arXiv:2505.18657*.