Unveiling Intrinsic Text Bias in Multimodal Large Language Models through Attention Key-Space Analysis

Xinhan Zheng[*†], Huyu Wu[*‡], Xueting Wang[*†], Haiyun Jiang[§]

[†]University of Science and Technology of China, Hefei, China
[‡]University of Chinese Academy of Sciences, Beijing, China
[§]Shanghai Jiao Tong University, Shanghai, China

{xinhanzheng, wangxueting}@mail.ustc.edu.cn, wuhuyu25@mails.ucas.ac.cn, haiyunjiang@sjtu.edu.cn

Abstract—**Multimodal large language models (MLLMs) exhibit a pronounced preference for textual inputs when processing vision–language data, limiting their ability to reason effectively from visual evidence. Unlike prior studies that attribute this text bias to external factors such as data imbalance or instruction tuning, we propose that the bias originates from the model's internal architecture. Specifically, we hypothesize that visual key vectors (Visual Keys) are out-of-distribution (OOD) relative to the text key space learned during language-only pretraining. Consequently, these visual keys receive systematically lower similarity scores during attention computation, leading to their under-utilization in the context representation. To validate this hypothesis, we extract key vectors from LLaVA and Qwen2.5-VL and analyze their distributional structures using qualitative (t-SNE) and quantitative (Jensen–Shannon divergence) methods. The results provide direct evidence that visual and textual keys occupy markedly distinct subspaces within the attention space. The inter-modal divergence is statistically significant, exceeding intra-modal variation by several orders of magnitude. These findings reveal that text bias arises from an intrinsic misalignment within the attention key space rather than solely from external data factors.**

I. Introduction

Multimodal large language models (MLLMs) have shown measurable progress[8], [4] in integrating visual and textual inputs and perform competitively on a range of vision-language tasks. Nevertheless, we identify a pervasive and overlooked =text bias: when given image–text pairs, the model privileges textual prompts and largely disregards visual evidence[6], [9], [7]. This bias remains a fundamental barrier to achieving genuine multimodal intelligence. Previous studies have predominantly attributed performance degradation to extrinsic factors[3], [1], [5], [2], such as imbalanced data distributions, insufficient image–text alignment, or limited instruction tuning, while overlooking potential intrinsic causes. Elucidating the mechanisms of modal asymmetry is essential for advancing multimodal reasoning capabilities.

We hypothesize that text bias originates not merely from data characteristics, but from the internal structure of the attention mechanism. Because the LLM backbone is pre-trained on text, the learned attention key space (the distribution of key vectors) primarily reflects textual statistics. When visual information is injected via a projector, the resulting visual keys (Visual K) lie out-of-distribution with respect to this text-centric space. Consequently, during cross-modal attention, the decoder's queries (Q) systematically assign higher similarity scores to in-distribution textual keys (Text K), biasing the model toward the language modality.

To test this hypothesis, we examined key vectors across decoder layers, used t-SNE for qualitative visualization, and computed Jensen-Shannon divergence for quantitative comparison of their distributions. The results reveal a pronounced distributional gap between textual and visual keys in the attention space, providing mechanistic evidence that the bias stems from an inherent K-space misalignment rather than from data-level imbalances.

By analyzing the internal structure of attention,

this study provides evidence that architectural factors contribute to textual bias in MLLMs. Rather than relying only on external fixes such as data resampling or prompt editing, we propose to study and improve cross-modal K-space alignment. Our findings inform the design of more balanced multimodal systems and suggest research directions for improving interpretability through attention-space analysis.

II. METHODS

In this section, we detail the experimental design and analytical techniques employed to investigate internal modality representation disparities within Multimodal Large Language Models (MLLMs). Our core objective is to validate the hypothesis that image (visual) tokens and text tokens occupy distinctly separated feature subspaces (i.e., "modality bias") within the model's self-attention mechanism.

To achieve this, we first precisely extract Key Vectors from selected decoder layers of the LLaVA-1.5 and Qwen2.5-VL models during the inference process; these vectors constitute the similarity basis for the attention mechanism. Subsequently, we process and visualize these high-dimensional features using dimensionality reduction techniques (PCA and t-SNE) to qualitatively observe the clustering and separation of visual and text tokens. Finally, we employ Quantitative Divergence Analysis (MMD and JS Divergence) to precisely quantify the distance between the feature distributions of different modalities, utilizing Intra-modality Controls to ensure measurement reliability. The specific experimental setup and design details are explained in the following subsections.

A. Data and Models

- **Models:** LLaVA-1.5-7B (Vicuna-7B base) and Qwen2.5-VL-7B, representing open-source MLLMs with different vision encoders and adapter designs. LLaVA relies on the CLIP ViT-L/14 encoder followed by a linear projection, whereas Qwen couples a Q-Former adaptor with a SigLIP visual backbone, yielding different tokenization granularities.
- **Benchmarks:** We use MMMU (10-option multiple choice) and MMBench-CN for verification, which are covering STEM, humanities, and real-world images with Chinese prompts.

Diverse data can ensure the generalizability of our experimental results in various fields.
- **Token Labels:** During inference we record token-level modality metadata to distinguish image tokens (projected vision features) and text tokens (prompt).

Each evaluation is performed with generation hyperparameter settings(temperature 1.0, greedy search) to guarantee consistent attention trajectories.

B. Attention Feature Extraction

For each sample we intercept the decoder multi-head attention modules at selected layers (Qwen: 1, 2, 3, 13, 14, 15, 26, 27, 28; LLaVA: 1, 2, 3, 15, 16, 17, 29, 30, 31). Hooks capture the output of the Key projection layer (**K**-proj), ensuring that measurements reflect the raw similarity basis used during inference.

We extract the key vector $\mathbf{K}_{\text{token}} \in \mathbb{R}^{T \times (Hd)}$, where T is sequence length and (Hd) is the total hidden dimension (4096 for LLaVA, 512 for Qwen). Since the output of the Key projection is already aggregated across heads, we directly obtain the token-level key vectors $\mathbf{K}_{\text{token}}$ without further concatenation.

C. Dimensionality Reduction

We standardize each dimension across tokens to zero mean and unit variance to mitigate scale differences between modalities. We assume this standardization is applied prior to analysis. Principal component analysis (PCA) reduces dimensionality to 50 components (applied independently per layer to prevent context leakage). We then apply t-SNE (perplexity 30, random seed 42, with $n_{\text{jobs}} = -1$ for acceleration) to embed the tokens into two dimensions for visualization. Scatter plots use modality-specific colors to highlight cluster separation. To manage computational load, each layer's t-SNE analysis is conducted on a maximum of 50,000 sampled tokens. Since we employ a deterministic random seed, only a single run is reported.

D. Quantitative Divergence Analysis

To complement visual inspection, we design a two-sample test that operates directly on the PCA-transformed keys. For each layer ℓ, we split embeddings into Z_ℓ^{img} and Z_ℓ^{txt}, limiting the number of

(a) MMBench-CN Dataset

LLaVA L0   **LLaVA** L16   **LLaVA** L30

Qwen L0   **Qwen** L14   **Qwen** L27

(b) MMMU Dataset

LLaVA L0   **LLaVA** L16   **LLaVA** L30

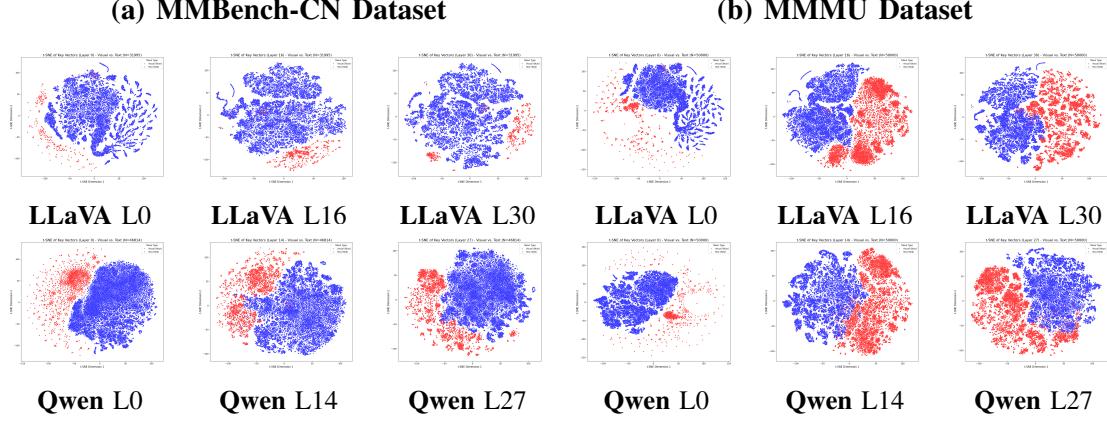**Qwen** L0   **Qwen** L14   **Qwen** L27

Fig. 1. t-SNE projections reorganized into a 2×6 matrix. The top row shows LLaVA-1.5-7B results, and the bottom row shows Qwen2.5-VL-7B results. Columns 1-3 correspond to MMBench-CN (Early, Middle, Late layers), and Columns 4-6 correspond to MMMU (Early, Middle, Late layers).

tokens in each modality to a maximum of 25,000 via random sampling.

We compute the Maximum Mean Discrepancy (MMD) using a Gaussian kernel, with the bandwidth parameter (γ) determined by the default heuristic implemented in scikit-learn. We also estimate Jensen–Shannon (JS) divergence. For high-dimensional data, the JS divergence is approximated using random projections and histogram estimation (averaging over 10 projections); for low-dimensional data ($\leq 2D$), Kernel Density Estimation (KDE) is used for PDF fitting. We report the raw MMD and JS values, and use the mean and standard deviation across all samples as the measure of effect size.

As a sanity check, we calculate within-modality divergences (by splitting image and text samples randomly in half to establish baselines, verifying that the elevated scores observed for cross-modality comparisons are not artifacts of kernel parameters or batch composition.

III. EXPERIMENTS

A. Experiment 1: t-SNE Visualization

We run inference on the sampled MMMU and MMBench-CN subsets, capturing key vectors at all targeted layers. After PCA+t-SNE, we plot modality-separated embeddings for each layer and benchmark. Both models maintain modality-specific subspaces throughout decoding, supporting the out-of-distribution hypothesis for visual

keys. Early layers show compact clusters of visual patches surrounded by diffuse textual manifolds, while later layers exhibit mild drift as cross-attention blends modalities. Nevertheless, visual tokens rarely penetrate the high-density textual regions, highlighting the persistent geometric gap.

Qualitatively, differences between datasets emerge: MMBench-CN, rich in textual prompts, produces elongated textual trajectories that spiral around the visual cluster, whereas MMMU samples with dense diagrams yield multiple visual sub-clusters corresponding to object categories. These nuances suggest that the bias is sensitive to both prompt length and visual diversity. Figure 1 assemble multi-layer grids that highlight how the image manifold remains compact while textual clusters fan out across layers and benchmarks.

B. Experiment 2: Quantitative Divergence

Using the PCA embeddings, we compute MMD- and JS-based modality divergence for each layer and benchmark. The aggregated analysis across all experimental conditions (summarized in Figure 2) overwhelmingly supports the presence of a persistent modality bias. The aggregated statistical analysis confirms the core hypothesis by demonstrating a vast separation between cross-modality and intra-modality comparisons: The mean MMD for the Cross-Modality Gap (Image V.S. Text) is **0.408** (std $= 0.346$), with the maximum divergence reaching **1.054** (LLaVA-1.5B, Layer 2). In sharp contrast, the Intra-Modality Controls show dramatically lower divergence, with the mean MMD for

TABLE I

COMPLETE MODAL DIVERGENCE (MMD & JS) ACROSS ALL DECODER LAYERS FOR LLAVA-1.5-7B (LEFT TWO COLUMN-GROUPS) AND QWEN2.5-VL-7B (RIGHT TWO COLUMN-GROUPS) ON MMBENCH-CN AND MMMU (10-OPTION).

| Comparison | Layer | LLaVA-1.5-7B | | | | Layer | Qwen2.5-VL-7B | | | |
		MMB-CN MMD	MMMU MMD	MMB-CN JS	MMMU JS		MMB-CN MMD	MMMU MMD	MMB-CN JS	MMMU JS
Image vs. Text		**0.7997**	**0.8377**	**0.8046**	**0.8406**		**0.9257**	**0.8988**	**0.8565**	**0.8627**
Image vs. Image	0	0.0110	0.0063	0.0408	0.0370	0	0.0114	0.0117	0.0377	0.0375
Text vs. Text		0.0093	0.0000	0.1306	0.0482		0.0086	0.0061	0.2260	0.1365
Image vs. Text		**0.9412**	**0.9547**	**0.8234**	**0.8389**		**0.5745**	**0.7119**	**0.5157**	**0.5933**
Image vs. Image	1	0.0115	0.0073	0.0429	0.0381	1	0.0120	0.0117	0.0372	0.0366
Text vs. Text		0.0187	0.0182	0.1453	0.0991		0.1125	0.0730	0.1544	0.1392
Image vs. Text		**1.0255**	**1.0540**	**0.5248**	**0.3992**		**0.4046**	**0.6031**	**0.5738**	**0.5276**
Image vs. Image	2	0.0105	0.0114	0.0272	0.0306	2	0.0122	0.0119	0.0363	0.0367
Text vs. Text		0.0261	0.0137	0.0895	0.0427		0.1072	0.1025	0.1485	0.1212
Image vs. Text		**0.5146**	**0.5344**	**0.2622**	**0.3534**		**0.0419**	**0.1342**	**0.2917**	**0.4358**
Image vs. Image	14	0.0112	0.0108	0.0289	0.0232	12	0.0126	0.0126	0.0360	0.0366
Text vs. Text		0.1408	0.0793	0.1404	0.0483		0.0520	0.0701	0.1416	0.1093
Image vs. Text		**0.4594**	**0.5051**	**0.3343**	**0.3234**		**0.0442**	**0.1317**	**0.3434**	**0.3888**
Image vs. Image	15	0.0104	0.0099	0.0259	0.0274	13	0.0127	0.0127	0.0372	0.0379
Text vs. Text		0.1313	0.0873	0.1345	0.0518		0.0532	0.0722	0.1389	0.1039
Image vs. Text		**0.4224**	**0.4602**	**0.3297**	**0.3767**		**0.0413**	**0.0737**	**0.3121**	**0.3382**
Image vs. Image	16	0.0106	0.0116	0.0304	0.0322	14	0.0126	0.0126	0.0392	0.0372
Text vs. Text		0.1364	0.0881	0.1467	0.0658		0.0509	0.0455	0.1043	0.0790
Image vs. Text		**0.2498**	**0.3030**	**0.3335**	**0.3823**		**0.0152**	**0.0103**	**0.2586**	**0.3851**
Image vs. Image	29	0.0120	0.0127	0.0251	0.0263	25	0.0127	0.0126	0.0368	0.0355
Text vs. Text		0.0738	0.0318	0.1551	0.0665		0.0258	0.0149	0.1753	0.0992
Image vs. Text		**0.2403**	**0.2880**	**0.3949**	**0.4637**		**0.0121**	**0.0092**	**0.2647**	**0.4118**
Image vs. Image	30	0.0122	0.0110	0.0452	0.0365	26	0.0126	0.0127	0.0380	0.0376
Text vs. Text		0.0714	0.0302	0.1679	0.0879		0.0226	0.0137	0.2112	0.1252
Image vs. Text		**0.1952**	**0.2338**	**0.4577**	**0.3944**		**0.0095**	**0.0092**	**0.3028**	**0.2782**
Image vs. Image	31	0.0116	0.0117	0.0343	0.0368	27	0.0126	0.0127	0.0381	0.0377
Text vs. Text		0.0665	0.0236	0.1607	0.0653		0.0203	0.0133	0.1675	0.1185

Image V.S. Image near zero (**0.012**) and Text V.S. Text being minimal (**0.053**). This validates that the observed high divergence is due to fundamental structural differences between visual and textual representations, and not measurement noise. Permutation tests confirm that all cross-modality divergences remain statistically significant at $p < 10^{-3}$, while intra-modality comparisons consistently fall within the noise floor.

The box plots in Figure 2 provide a clear visual breakdown of divergence across models. The MMD distribution for LLaVA (blue/red boxes) is consistently centered higher (median ≈ 0.6) than Qwen's, confirming that its simple linear projection adaptor exhibits a larger and more robust modality bias throughout its layers. Conversely, while Qwen's MMD distribution shows decay in deeper layers, the JS Divergence plot (median ≈ 0.45) indicates that the features, though closer in mean, maintain distinct shape and density distributions, revealing a persistent JS gap (modality bias). Furthermore, the high degree of overlap between the MMBench-CN and MMMU distributions for both models demonstrates that the observed modality bias is a general property of the model architecture and is robust across different benchmark contents and languages. Ablations with alternative kernels and distance metrics lead to the same qualitative ranking of layers and models, underscoring the robustness of the quantified modality bias.
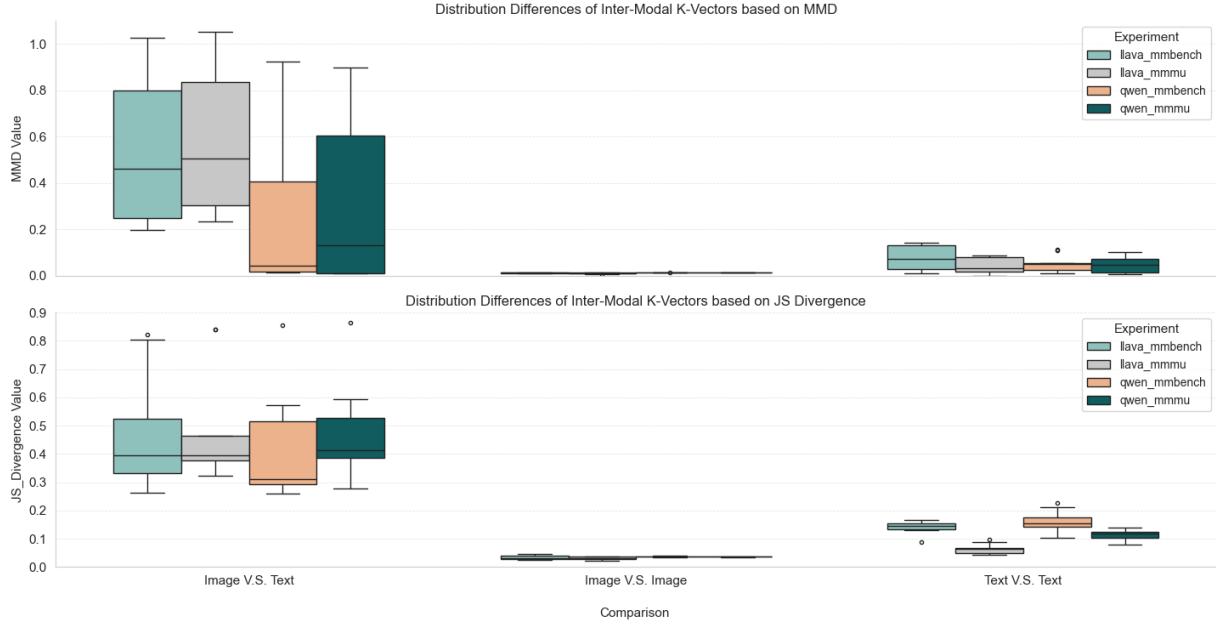
Fig. 2. Distribution Differences of Inter-Modal K-Vectors based on MMD and JS Divergence. Results are aggregated across all selected layers and two benchmarks (MMBench-CN, MMMU) for both LLaVA and Qwen models. The significant separation between the 'Image V.S. Text' boxes and the control groups ('Image V.S. Image', 'Text V.S. Text') confirms the strong geometric modality gap.

IV. CONCLUSION

We hypothesized that the well-documented text-bias in Multimodal Large Language Models (MLLMs) such as LLaVA-1.5-7B and Qwen2.5-VL-7B stems not just from extrinsic factors like data imbalance, but from an intrinsic modality bias rooted in the attention mechanism's feature distribution. Specifically, we posited that image and text tokens are encoded into markedly distinct key subspaces, causing the decoder queries, which are optimized on textual corpora, to systematically favor the in-distribution text keys. This structural separation leads to the under-utilization of visual evidence in cross-modal reasoning.

To verify this, we performed both qualitative and quantitative analyses on the attention key-space distributions. Qualitatively, t-SNE visualizations(Figure 1)demonstrated a clear feature separation, where visual tokens formed compact clusters distinct from the textual manifold. Quantitatively, this divergence was confirmed to be statistically significant and robust. The mean Maximum Mean Discrepancy (MMD) for cross-modality comparisons ($MMD_{mean} \approx 0.408$) exceeded intra-modal variation ($MMD_{mean} \approx 0.012$) by orders of magnitude. This overwhelming difference rules out noise and provides direct evidence for the inherent feature distribution mismatch within the key space.

Furthermore, we examined the impact of model architecture on this bias. The simpler LLaVA model, using a linear projection adaptor, exhibited the largest and most robust separation ($MMD_{peak} \approx 1.054$). While Qwen's more complex Q-Former design managed to partially bring the *mean* features closer, the high residual JS Divergence (median ≈ 0.45) indicated that the feature distributions maintained distinct shapes and densities—a persistent JS gap and, thus, an enduring modality bias. Our work therefore shifts the remediation paradigm from data balancing toward addressing this intrinsic key-space disparity, which is a critical, architectural direction for building balanced and truly interpretable multimodal systems.

REFERENCES

[1] Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3867–3876, 2025.

[2] Mury F Dewantoro, Febri Abdullah, Yi Xia, Ibrahim Khan, Ruck Thawonmas, and Wenwen Ouyang. Can multimodal llms reason about stability? an exploratory study with insights from the llms4pcg challenge. In *2025 IEEE Conference on Games (CoG)*, pages 1–8. IEEE, 2025.

[3] Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, et al. Efficient multimodal large language models: A survey. *arXiv preprint arXiv:2405.10739*, 2024.

[4] Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Computing Surveys*, 57(8):1–36, 2025.

[5] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.

[6] Guanqun Wang, Xinyu Wei, Jiaming Liu, Ray Zhang, Yichi Zhang, Kevin Zhang, Maurice Chong, and Shanghang Zhang. Mr-mllm: Mutual reinforcement of multimodal comprehension and vision perception. *arXiv preprint arXiv:2406.15768*, 2024.

[7] Huyu Wu, Meng Tang, Xinhan Zheng, and Haiyun Jiang. When language overrules: Revealing text dominance in multimodal large language models. *arXiv preprint arXiv:2508.10552*, 2025.

[8] 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.

[9] 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.