Unveiling Bias in Multimodal Models

by
Yogesh Prabhu

A thesis submitted in partial fulfillment of
the requirements for the degree of
Bachelor of Science, Honors in the Major
(Computer Science)

University of Wisconsin Madison
Department of Computer, Data & Information Sciences

Supervisor:
Prof. Frederic Sala

Abstract

Vision Language Models (VLMs) have significantly advanced multimodal understanding by effectively combining visual and textual modalities for various applications, including image captioning, visual question answering, and video summarization. However, despite their capabilities, these models exhibit pronounced modality biases, predominantly relying on textual inputs over visual data. This thesis systematically evaluates unimodal biases in state-of-the-art VLMs, highlighting the impacts on performance and proposing innovative strategies for bias mitigation, including Prompting, Interleaved Vision-Text Projection (IVTP), and Cross-Attention Projection. Our experimental evaluations using the MMWorld dataset demonstrate that targeted mitigation strategies substantially enhance modality balance and model robustness. The findings underscore the importance of architectural adjustments and training methodologies to ensure equitable multimodal integration, paving the way for more reliable and robust multimodal AI systems.

Acknowledgments

I would like to extend my heartfelt gratitude to my supervisor, Prof. Frederic Sala, for his unwavering support, guidance, and encouragement throughout the entire journey of this thesis. His valuable insights, constructive feedback, critical review, and patience have played a pivotal role in shaping this research.

I am deeply thankful to the L&S Honors Program for granting me the incredible opportunity to complete this thesis and for the privilege of presenting my work at the Senior Honors Thesis Symposium 2024. I am also grateful to the University of Wisconsin-Madison for providing access to all necessary research resources and to the Computer Science Department for fostering an exceptional academic environment. The dedication and expertise of the department's faculty and staff have greatly enriched my undergraduate experience.

To my family, I express my profound appreciation for their love, encouragement, and steadfast support throughout this journey. Their sacrifices and belief in my aspirations have been the driving force behind my achievements.

A special note of thanks to PhD student Dyah Adila for her invaluable guidance and assistance with this project.

This thesis is the result of the collective contributions of everyone mentioned above. Thank you all for believing in me and for being an integral part of this experience.

Contents

1 Introduction

Multimodal machine learning models have demonstrated extraordinary capabilities in understanding, integrating, and generating information across different modalities, particularly image, text, and video. Vision Language Models (VLMs), which integrate visual and textual information [32, 3], have emerged at the forefront of this field, enabling impressive performance across diverse tasks such as image captioning, visual question answering, and video summarization [43, 2, 33]. However, alongside their rapid progress and adoption, concerns have emerged regarding modality bias—specifically, the tendency of multimodal models to disproportionately rely on one modality (usually text) rather than evenly integrating information from multiple modalities. This modality bias compromises the intended balanced interaction across modalities, impacting the overall effectiveness and robustness of these models.

1.1 Background

Recent years have witnessed extraordinary advancements in natural language processing, particularly with the emergence of large language models (LLMs) such as GPT-4 [31], LLaMA [41], and Gemma[39]. Following these breakthroughs in text understanding and generation, researchers have increasingly focused on integrating additional modalities into these language models, seeking to develop systems that can process and reason across different forms of information simultaneously.

These multimodal integration efforts have followed several distinct approaches. Some researchers have pursued the strategy of connecting separately pre-trained unimodal models, where vision models and language models are first trained independently on their respective modalities before being integrated through specialized alignment layers [25]. Others have advocated for training multimodal architectures from scratch on paired multimodal data, allowing for more organic integration but requiring substantially more computational resources. More recently, hybrid approaches have emerged that start with a strong language model foundation and extend it to process visual inputs through specialized adapters or projection mechanisms [44].

These advancements in artificial intelligence, particularly in multimodal learning, have facilitated the emergence of sophisticated Vision Language Models. Prominent examples such as VideoLLaMA 3, Qwen2.5-VL, Gemma 3, and LLaMA 3.2 illustrate significant strides achieved in capturing and correlating visual, textual, and video modalities. However, recent analyses suggest that despite their multimodal design, these models often exhibit "unimodal bias," where they overly rely on textual information at the expense of visual or video content [7]. This bias arises because textual modality typically offers explicit, readily accessible semantic cues, whereas extracting nuanced semantic information from visual or temporal content remains more challenging.

Unimodal bias may lead to performance degradation in tasks where visual or video information is critical, especially when textual context is ambiguous or insufficient [21]. Thus, understanding and addressing modality bias is crucial to ensure that multimodal models effectively leverage all available information.

1.2 Problem Statement

Although Vision Language Models are explicitly designed to integrate multiple modalities, we find that they frequently exhibit a bias toward textual modality, undermining the balanced utilization of visual and video information. Such modality bias diminishes model performance on tasks requiring holistic multimodal understanding and risks creating misleading outcomes when textual cues are unclear or insufficiently descriptive. This thesis aims to specifically investigate:

- The extent and conditions under which modality bias occurs in state-of-the-art multimodal models.
- The consequences of modality bias on performance and reliability in tasks requiring genuine multimodal integration.
- Strategies to mitigate modality bias, fostering a balanced and effective multimodal interaction.

1.3 Objective

The primary objective of this research is to systematically analyze, quantify, and address unimodal (specifically textual) bias within state-of-the-art Vision Language Models. Specific aims include:

1. Developing robust methodologies for detecting and quantifying modality bias, focusing particularly on biases favoring textual modality.
2. Evaluating the impact of modality bias on model performance across diverse multimodal scenarios.
3. Proposing effective techniques and architectures to promote balanced multimodal integration and reduce over-reliance on textual information.

Through achieving these objectives, the research will contribute to advancing the fairness, robustness, and multimodal capabilities of Vision Language Models.

1.4 Thesis Outline

The thesis is structured as follows:

- Chapter 2: Literature Review – This chapter presents a comprehensive review of related work on modality bias, multimodal learning techniques, and existing methods for detecting and addressing modality bias in AI systems.
- Chapter 3: Methodology – This chapter details our proposed approaches for detecting, quantifying, and mitigating modality bias specific to multimodal Vision Language Models.
- Chapter 4: Experiments and Results – This chapter provides detailed experimentation demonstrating modality bias, including quantitative analyses and case studies highlighting bias in practical scenarios.
- Chapter 5: Discussion – This chapter offers interpretation of results, implications of modality bias, and exploration of the limitations and strengths of proposed solutions.
- Chapter 6: Conclusion and Future Directions – This chapter summarizes our contributions and suggests directions for further exploration in multimodal bias research.

2 Literature Review

This chapter reviews relevant literature on multimodal models, bias in AI, and specifically addresses the issue of unimodal bias in multimodal systems. We explore state-of-the-art multimodal models, discuss bias in AI broadly, and provide an in-depth analysis of unimodal bias.

2.1 Multimodal Models

Multimodal machine learning integrates multiple data types—such as text, images, audio, and videos—to perform complex tasks requiring simultaneous comprehension across diverse domains. Recent developments have yielded advanced Vision Language Models (VLMs) capable of sophisticated multimodal understanding and content generation. This integration aims to mirror human cognitive abilities to process information through multiple sensory channels simultaneously.

Current state-of-the-art multimodal models include VideoLLaMA 3, Qwen2.5-VL, Gemma 3, and LLaMA 3.2, each demonstrating remarkable capabilities in text-image and text-video tasks. VideoLLaMA 3 utilizes variable-length visual tokens, effectively managing varying image and video resolutions, thereby excelling in both static and dynamic visual comprehension [49]. Qwen2.5-VL enhances multimodal functionality through dynamic image resolution handling and temporal encoding, optimizing performance on long-form visual content [5]. Gemma 3, developed by Google, is a lightweight, multilingual multimodal model optimized for environments with limited computational resources, yet still robust in multimodal reasoning tasks [38]. LLaMA 3.2 incorporates visual capabilities into Meta AI's open-source language model family, significantly expanding its utility in multimodal scenarios [13].

These models typically employ transformer-based architectures [42], extensive pre-training on large multimodal datasets, and advanced fine-tuning techniques. Despite their impressive capabilities, they are susceptible to modality-specific biases resulting from challenges inherent in integrating heterogeneous data types.

The architectures of multimodal large language models (MLLMs) vary significantly, generally categorized into three primary types:

1. **Unified Embedding Decoder Architecture**: This approach involves projecting video or image embeddings generated by an image encoder into the embedding space of an LLM, typically using a multilayer perceptron (MLP). Models such as VideoLLaMA 3, Qwen2.5-VL, and Gemma 3 follow this unified embedding approach [38, 5, 49]. While efficient in terms of parameter usage, this design can potentially favor the dominant modality in the embedding space, often text, due to the unidirectional

projection mechanism.

2. **Cross-Modality Attention**: Exemplified by LLaMA 3.2, this architecture employs attention mechanisms that allow visual embeddings derived from an image encoder to interact with textual representations through dedicated cross-attention layers [13]. This architecture facilitates dynamic bidirectional interactions between textual and visual representations, enabling richer multimodal integration and understanding. The distinct advantage of this approach is its ability to maintain separate representational spaces for each modality before fusion, potentially mitigating modality dominance.

3. **Vision as LoRA (VoRA)**: A recent architectural innovation that, unlike traditional methods relying on external vision modules, integrates vision-specific LoRA (Low-Rank Adaptation) layers directly within the LLM itself [45]. This structural modification enables seamless merging of visual capabilities into the language model during inference, reducing computational overhead and simplifying the model architecture. VoRA's design permits handling inputs at arbitrary resolutions due to its intrinsic adaptability. VoRA incorporates a novel block-wise distillation method that transfers visual priors from a pre-trained Vision Transformer (ViT) [11] into the LoRA layers, significantly expediting training [45]. Furthermore, bi-directional attention masks enhance the model's capacity to capture comprehensive contextual information from images. Empirical evaluations have demonstrated that with adequate pre-training, VoRA achieves comparable performance to traditional encoder-based MLLM architectures while requiring fewer parameters.

Beyond these architectural approaches, recent research has explored various fusion strategies at different levels of model depth. Nagrani et al.[30] investigated fusion bottlenecks for multimodal integration, demonstrating that restricting cross-modal attention through bottleneck tokens can improve performance while reducing computational complexity. Their work showed that the placement of fusion layers within the network hierarchy significantly impacts how effectively modalities are integrated, with mid-level fusion often outperforming both early and late fusion strategies.

2.2 Bias in AI Models

Biases in AI models refer to systematic errors or preferences embedded within a model, often arising from training data, architectural design, or specific task formulations. Such biases can negatively affect model fairness, robustness, and generalizability, potentially leading to harmful outcomes, especially in sensitive applications.

AI biases generally fall into three categories:

Data Biases: Models trained on imbalanced or modality-specific datasets can

exhibit biases in how they process and prioritize different types of input. This is especially relevant in the case of multimodal models, where a common approach is to use a large pretrained language model (LLM) as the backbone. These LLMs are typically trained primarily on vast amounts of text data before being extended to handle other modalities such as images. As a result, the models may develop a modality bias—favoring textual information over visual cues—since the foundational representations are rooted in the text modality.

The issue is compounded by the fact that multimodal datasets often contain inherent correlations between modalities that may not reflect causal relationships. For example, images of snowy scenes are frequently paired with text mentioning winter, creating a spurious correlation that models may exploit rather than learning deeper semantic connections. These dataset biases can lead models to make predictions based on superficial patterns rather than meaningful multimodal understanding.

Architectural Biases: Biases may also arise from model architectures or training methodologies. Models can develop unintended dependencies on dominant modalities or simplistic heuristics, undermining their intended balanced performance across different input types [6, 50]. For example, the Unified Embedding Decoder architecture might inherently favor textual information due to direct embedding projection, potentially introducing a bias towards the textual modality. In contrast, architectures employing Cross-Modality Attention mechanisms, such as LLaMA 3.2, can help mitigate such biases by allowing dynamic interaction between modalities, thus enabling the model to better balance its attention between textual and visual inputs.

The gradient flow during training also significantly influences modality bias. Zhang et al.[50] theoretically demonstrated that the convergence of loss functions in multimodal networks is heavily influenced by both the magnitude and direction of gradients. When gradients from different modalities conflict or vary substantially in magnitude, the stronger modality dominates the learning process, leading to an unbalanced model that relies primarily on that modality.

Task Design Biases: Tasks designed without balanced modality demands can inadvertently encourage unimodal biases. If one modality consistently provides stronger signals than others, models may learn to ignore less predictive modalities entirely, resulting in incomplete or skewed multimodal integration [1, 14]. For instance, in visual question answering tasks, if questions consistently contain strong linguistic priors (e.g., questions about color almost always referring to prominent objects), models may learn to answer based on text patterns alone, effectively ignoring the visual input.

Benchmark design plays a crucial role in either mitigating or reinforcing these biases. Traditional VQA datasets were found to contain strong language priors that

allowed models to perform well without truly understanding images. This led to the development of more challenging benchmarks like VQA-CP (Visual Question Answering with Changing Priors), which intentionally alters the distribution of answers between training and testing to discourage reliance on linguistic shortcuts.

Thus, understanding the nature and source of biases is crucial for developing reliable multimodal AI systems. Recent work has focused on developing metrics to quantify these biases, creating balanced evaluation datasets, and designing training strategies that explicitly counter modality imbalance.

2.3 Unimodal Bias in Multimodal Learning

Unimodal bias is a specific form of bias occurring when multimodal models disproportionately rely on one modality, often text, rather than equally leveraging all available modalities. This tendency diminishes the advantages of multimodal learning, limiting the model's robustness and effectiveness in tasks genuinely requiring multimodal reasoning.

Zhang et al.[50] provide a comprehensive theoretical framework for understanding unimodal bias, particularly in multimodal deep linear networks. Their analysis identifies how late-stage fusion in architectures exacerbates the duration of a "unimodal learning phase," during which models primarily exploit easier-to-learn modalities, such as text, neglecting visual or temporal modalities. Their work mathematically demonstrates that the deeper the fusion layer within a network, the longer the unimodal phase persists, leading to what they term "permanent unimodal bias" in overparameterized models. This theoretical understanding highlights the importance of architectural choices in mitigating unimodal bias.

The mechanism underlying unimodal bias can be understood through the lens of learning dynamics. When training multimodal models, modalities with stronger correlation to target outputs tend to dominate early learning stages. This creates a "Matthew effect" where the stronger modality continues to strengthen while weaker modalities are neglected. Zhang et al.[50] show that this effect is particularly pronounced in late fusion architectures where modalities are processed independently until the final layers. In contrast, early fusion architectures, which combine modalities at input stages, show more balanced learning dynamics but may struggle with heterogeneous feature extraction.

Empirical studies further confirm the prevalence of unimodal bias. For instance, models evaluated on the Visual Question Answering (VQA) benchmark frequently exploit linguistic priors instead of genuinely integrating visual content, failing when faced with counterfactual or modified questions [1, 6]. Similarly, multimodal hate speech detection tasks, such as the Hateful Memes Challenge, revealed that models commonly over-rely on

textual inputs, overlooking crucial visual context needed for accurate classification [18].

Kwon et al.[20] identified through experimental analysis that certain modalities are more crucial to target performance than others, varying by task and dataset. Their work demonstrates that dominant and weak modalities converge at different rates during training, with the dominant modality's loss decreasing rapidly while the weak modality's loss decreases more slowly. They theoretically prove that this imbalance results from unaligned gradients or differences in gradient magnitudes that prevent balanced convergence of the loss function.

The practical implications of unimodal bias are significant, particularly in safety-critical applications. A multimodal system biased toward textual inputs might fail to recognize visual danger signals or misunderstand situations where visual and textual cues provide contradictory information. In medical diagnosis, for example, a model overly reliant on textual patient history might miss crucial visual indicators in medical imaging. Similarly, autonomous vehicles heavily biased toward one sensory modality could make dangerous decisions when that modality is impaired.

Mitigation strategies for unimodal bias include architectural adjustments, specialized training methods, and enhanced evaluation benchmarks. Techniques such as auxiliary classifiers (e.g., RUBi) that penalize unimodal predictions, gradient balancing frameworks (e.g., BalGrad), and attention bottlenecks have been proposed to enforce more balanced multimodal integration [6, 20, 30].

Nagrani et al.[30] introduced a novel approach using "fusion bottlenecks" that restrict the flow of cross-modal information between latent units through tight bottleneck tokens. This forces the model to collect and condense the most relevant information from each modality before sharing it with others, improving fusion performance while reducing computational cost. Their experiments on audio-visual classification tasks demonstrated that this approach outperforms both early and late fusion strategies.

Kwon et al.[20] proposed BALGRAD, a framework addressing dominant modality bias through two key components: inter-modality gradient reweighting and inter-task gradient projection. The former adjusts the magnitude of the gradient based on each modality's learning status, while the latter aligns task directions in a non-conflicting manner. Their approach effectively reduced the performance gap between modalities across multiple datasets.

Additionally, carefully designed benchmarks like VQA-CP and Hateful Memes explicitly discourage unimodal solutions, prompting models toward genuine multimodal learning [1, 18].

Despite these advances, unimodal bias remains a challenging and critical issue

in multimodal AI, demanding further research into effective detection and mitigation strategies, particularly as models scale to handle increasingly complex multimodal tasks [8].

2.4 Optimal-Transport Perspectives on Multimodal Alignment

Optimal Transport (OT) offers a principled way to quantify how costly it is to morph one probability distribution into another [28]. Classic results define, for two measures α, β over a metric space $\mathcal{X}$, the p-Wasserstein distance [12]

$$W_p(\alpha, \beta) \;=\; \Big(\min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{X}\times\mathcal{X}} d(x, x')^p \, \mathrm{d}\pi(x, x')\Big)^{1/p},$$

where $\Pi(\alpha, \beta)$ is the set of couplings having α and β as marginals . Recent work adapts this idea to datasets whose samples carry both features and labels. Alvarez-Melis & Fusi [4] introduce the Optimal-Transport Dataset Distance (OTDD): labels are mapped to class-conditional feature distributions, and an OT problem is solved in the product space $\mathcal{X} \times \mathcal{P}(\mathcal{X})$ using a compound cost

$$d((x, y), (x', y')) = \big(d_{\mathcal{X}}(x, x')^p + W_p(P_{X|Y=y},\, P_{X|Y=y'})^p\big)^{1/p}.$$

The resulting distance reflects both geometric proximity of features and semantic congruence of labels, properties that are essential when aligning heterogeneous modalities.

In multimodal learning, OT has been adopted in three complementary roles:

- **Distribution matching for pre-training**: Works such as ORCA [36] first learn a lightweight embedder that minimises OTDD between embedded target data and an in-modality reference set before fine-tuning a frozen transformer backbone. Empirically, OTDD outperforms alternatives like Maximum Mean Discrepancy and Euclidean matching across twelve modalities.

- **Regularisation during joint optimisation**: Instead of a pre-alignment stage, some vision–language architectures inject a Wasserstein term directly into the training loss to penalise divergence between visual and textual latent distributions, encouraging a smoother shared space and mitigating "posterior collapse".

- **Post-hoc calibration**: OT barycentres can fuse predictions from separate uni-modal experts into a single consensus distribution that respects modality-specific uncertainties.

Current challenges include the cubic complexity of exact OT, choice of ground metric in high-dimensional latent spaces, and extending OTDD beyond discrete labels (video and speech often entail continuous targets). The entropy-regularised Sinkhorn

solver and class-wise subsampling proposed in ORCA reduce computation from $\mathcal{O}(n^3)$ to roughly $\mathcal{O}(n^2)$ without hurting downstream accuracy.

2.5 Open Challenges

Existing literature underscores both the progress and persistent challenges associated with unimodal bias in multimodal models. Current state-of-the-art approaches highlight the complexity of balancing multiple modalities effectively, particularly under realistic conditions where modalities vary significantly in informativeness.

Key open challenges identified include:

1. **Scalable Measurement Frameworks**: While various methods exist to detect unimodal bias, developing standardized, computationally efficient metrics that scale to large models and diverse tasks remains challenging. Current approaches often require task-specific adaptations or extensive computational resources for ablation studies.

2. **Adaptive Mitigation Strategies**: Current methods for mitigating unimodal bias often require manual tuning or task-specific adjustments. Developing self-adapting approaches that automatically calibrate modality integration based on task demands and modality informativeness would significantly advance the field.

3. **Balancing Bias Mitigation with Performance**: A critical tension exists between enforcing balanced modality usage and maintaining optimal task performance [37]. In some cases, certain modalities genuinely contain more task-relevant information. Determining when to enforce balance versus when to allow natural modality dominance remains an open question.

4. **Addressing Temporal Dynamics in Video Understanding**: Most current work focuses on static multimodal fusion (image-text pairs). The additional complexity of temporally evolving modalities in video understanding introduces new challenges for balanced integration that are underexplored.

5. **Transfer Learning Across Modality Configurations**: Developing models that can robustly transfer knowledge across different modality combinations and varying levels of modality availability would enhance real-world applicability [46].

6. **Beyond Binary Modality Comparisons**: Much existing research focuses on binary combinations (typically vision-language). Extending bias mitigation approaches to genuinely multi-modal settings with three or more modalities introduces additional complexity that remains largely unexplored.

7. **Explainable Multimodal Integration**: Developing techniques to not only bal-

ance modality usage but also explain how and why different modalities contribute to decisions would enhance trust and enable more targeted improvements [10].

8. **Real-world Robustness Testing**: Evaluating unimodal bias under authentic conditions where modalities may be impaired, noisy, or partially unavailable represents a crucial gap in current assessment methodologies.

This thesis addresses these challenges by proposing novel methodologies for analyzing, quantifying, and mitigating unimodal biases, contributing toward more equitable and robust multimodal AI systems. Our work specifically focuses on developing adaptive mitigation strategies that can balance modality usage without compromising performance, introducing new metrics for quantifying unimodal bias at scale, and evaluating these approaches across diverse real-world scenarios.

3 Methodology

This section presents our methodological framework for evaluating and mitigating unimodal bias in multimodal Vision-Language Models (VLMs). We first establish a rigorous evaluation protocol using the MMWorld dataset [15], then detail our novel approaches to address modality imbalance through architectural and training innovations. Our methodology is designed to systematically quantify bias patterns and implement targeted interventions that promote balanced information integration across modalities.

3.1 Evaluation Framework

3.1.1 Dataset Characteristics

We utilize the MMWorld dataset [15], which comprises 12,500 video-text pairs spanning 7 discipline categories. Each sample contains: a 10-second video clip at 30fps with 1080p resolution, a detailed caption averaging 15.3 words, and a ground-truth classification label. The dataset exhibits balanced class distribution (500 samples per category) and was manually curated to ensure high inter-annotator agreement ($\kappa > 0.92$).

3.1.2 Experimental Conditions

To systematically assess unimodal bias, we design a comprehensive set of experimental conditions that strategically manipulate the relationship between modalities:

1. **Baseline (BL)**: Both video and caption correctly correspond to the classification label, serving as our control condition to establish optimal multimodal integration performance.

2. **Text Bias Assessment (TB)**: Videos are completely replaced with blank frames (uniform gray) while preserving accurate captions, isolating text-only processing capabilities.

3. **Visual Bias Assessment (VB)**: Captions are entirely removed, leaving only correctly labeled videos to evaluate pure visual processing.

4. **Adversarial Modality Conflict (AMC)**: We deliberately introduce conflicting information across modalities:

    - **AMC-V**: Videos from category A are paired with captions from category B that contain strong semantic contradictions.

    - **AMC-T**: Captions from category A are paired with videos from category B that contain strong semantic contradictions.

3.1.3 Metrics and Analysis

We employ a multi-dimensional evaluation approach to quantify unimodal bias:

$$\text{Text Bias Index (TBI)} = \frac{\text{Acc}_{\text{TB}} - \text{Acc}_{\text{AMC-T}}}{\text{Acc}_{\text{BL}}} \tag{1}$$

$$\text{Visual Bias Index (VBI)} = \frac{\text{Acc}_{\text{VB}} - \text{Acc}_{\text{AMC-V}}}{\text{Acc}_{\text{BL}}} \tag{2}$$

$$\text{Modality Reliance Score (MRS)} = \max\left(\text{Acc}_{\text{AMC-T}}, \text{Acc}_{\text{AMC-V}}\right) - \min\left(\text{Acc}_{\text{AMC-T}}, \text{Acc}_{\text{AMC-V}}\right) \tag{3}$$

The TBI and VBI quantify relative bias toward each modality, while MRS measures absolute disparity in multimodal processing capabilities. Ideally we would want to see MRS as close to 0 as possible, and we would want to see TBI and VBI be simillar in values while both being close to 0.

3.2 Bias Mitigation Approaches

We propose three complementary approaches to mitigate unimodal bias, each targeting different aspects of multimodal integration:

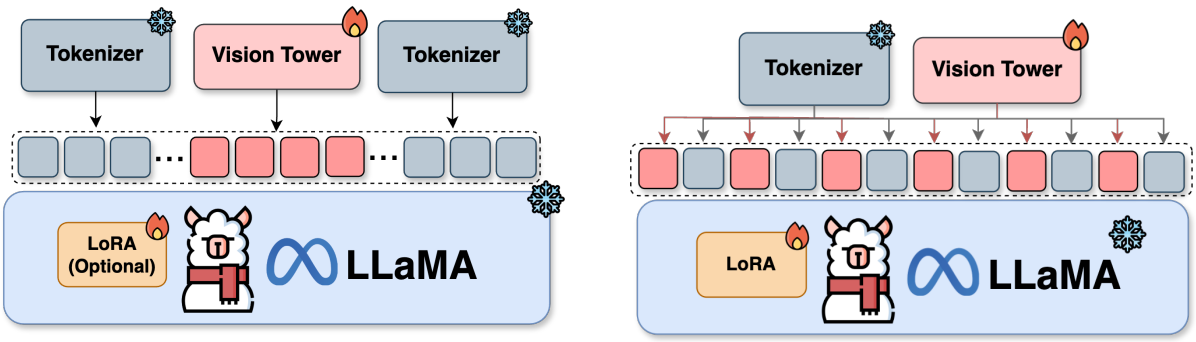
3.2.1 Prompting

Prompt-based strategies involve explicitly instructing models to consider each modality equally. By integrating explicit instructions into prompts—for example, emphasizing "please pay equal attention to both textual and visual content"—we encourage models to balance modality usage [34]. This method leverages the inherent linguistic capabilities of language models to adjust their attention dynamically during inference [34]. We systematically evaluate a range of prompt templates that emphasize balanced modality integration. Example prompts include:

- "Consider both the video and text equally in your analysis."
- "The video and text may contain different information; evaluate both carefully."
- "Weigh visual and textual information equally in making your determination."

This approach requires no architectural modifications, making it easily applicable across different model configurations while potentially addressing inherent modality biases through explicit guidance.

While prompting offers an efficient implementation without model retraining, we anticipate that controlling unimodal bias through this method alone will have limited effectiveness. This is because exploring the complex discrete space of prompts is challenging, and deeply embedded biases in model weights are unlikely to be fully overcome



(a) Qwen2.5-VL approach [5]: text tokens (grey) at beginning and end, vision tokens (red) in the middle.

(b) IVTP approach: interleaves vision tokens with text tokens in an alternating pattern.

Figure 1: **Interleaved Vision-Text Projection architecture.** Comparison between the Qwen2.5-VL approach (left) and our IVTP approach (right).

through instruction alone. We expect modest improvements in modality balance, but significant bias may persist, particularly in cases with strong modality conflicts.

3.2.2 Interleaved Vision-Text Projection (IVTP)

The IVTP approach fundamentally restructures the input representation to enforce balanced multimodal integration. As illustrated in Figure 2, the standard approach processes vision and text tokens sequentially, which can lead to modality neglect as the model may prioritize one modality over another. Our IVTP implementation addresses this by systematically interleaving vision tokens between text tokens in an alternating pattern. The key components of our implementation include:

1. **Tokenizer and Vision Tower Integration**: Inspired by Qwen2.5-VL [5] We maintain separate encoders for text (Tokenizer) and visual input (Vision Tower), but modify how their outputs are combined. The Vision Tower processes images to produce visual embeddings in between text while the Tokenizer converts the prompt into token embeddings and a prefix into embeddings before the visual embeddings.

2. **Alternating Token Arrangement**: Rather than keeping vision tokens as a contiguous block, we distribute them evenly throughout the sequence, creating a regular alternating pattern of text and vision tokens [40]. This forces the model to continuously switch between modalities during processing.

3. **Parameter-Efficient Fine-tuning**: We implement this interleaved architecture using LoRA (Low-Rank Adaptation) [16] to efficiently adapt the pre-trained LLaMA model [47] for the interleaved token processing, minimizing additional parameter overhead while enabling effective cross-modal learning.

This architecture enforces balanced processing by ensuring that visual information is

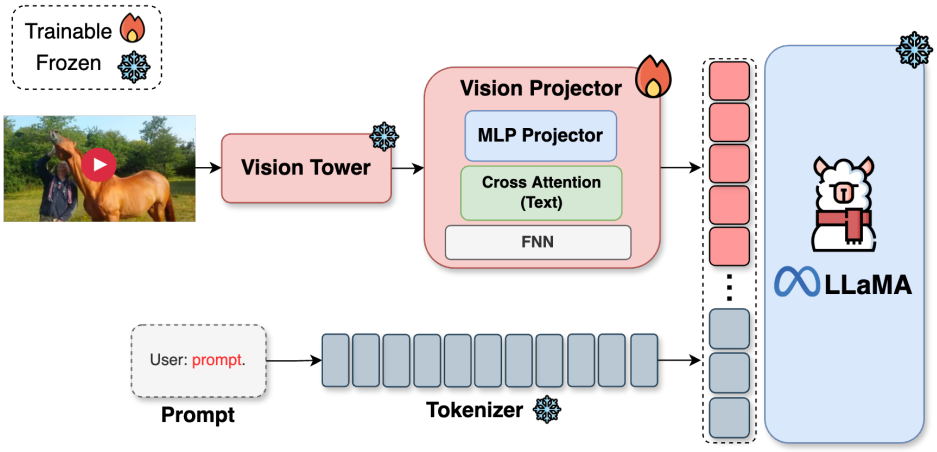


Figure 2: **Cross-Attention Projector architecture.** The approach employs a Vision Tower to extract visual features, which are then processed through a Vision Projector containing multiple components: an MLP Projector for dimensionality alignment, Cross-Attention layers that condition visual features on text embeddings [29], and a Feed-Forward Network (FNN) for final transformation. The processed visual tokens are then concatenated with text tokens before being fed into the LLaMA model.

consistently integrated throughout the sequence, preventing the model from focusing exclusively on a single modality during the autoregressive decoding process.

We expect the IVTP approach to yield more substantial improvements in modality balance compared to prompting, as it directly addresses the structural issue of modality segregation in the input sequence. By forcing the model to repeatedly alternate between modalities, we anticipate reduced opportunities for the model to "ignore" one modality in favor of another. However, we also expect that this approach may increase computational complexity and potentially reduce performance on tasks where one modality genuinely contains more relevant information than the other.

3.2.3 Cross-Attention Projector

The Cross-Attention Projector architecture provides a sophisticated mechanism for integrating visual information with textual data as shown in Figure 2. This approach consists of the following key components:

1. **Vision Tower**: The video input is processed by a pre-trained Vision Tower (with frozen weights) to extract high-quality visual features. This component transforms raw visual input into meaningful representations.

2. **Vision Projector**: A trainable component that processes the visual features through three sub-modules:

    - **MLP Projector**: Aligns the dimensionality of visual features with the lan-

guage model's embedding space.

- **Cross-Attention (Text)**: Conditions the visual features on textual context, enabling the visual tokens to be informed by and adapt to relevant textual information [29].
- **Feed-Forward Network (FNN)**: Further processes the cross-attended visual representations to produce the final visual tokens.

3. **Multimodal Integration**: The processed visual tokens are prepended to the text tokens derived from the prompt through the Tokenizer. This arrangement allows the LLaMA model to first process the visual information before addressing the textual content, potentially mitigating text bias.

This architecture provides a more balanced approach to modality integration by explicitly using text information to guide the processing of visual features through cross-attention, ensuring that visual information is meaningfully incorporated rather than ignored during inference.

We hypothesize that the Cross-Attention Projector will provide the most effective mitigation of unimodal bias among our proposed approaches, as it not only ensures structural integration of modalities but also enables dynamic, content-dependent interactions between them. The bidirectional flow of information through cross-attention should allow the model to weigh modalities more appropriately based on their informational content for the specific task. However, we also anticipate that this approach will be the most computationally intensive and may require more extensive fine-tuning to realize its full potential.

3.3 Implementation Details

All experiments were conducted using the Panda-70M dataset [9] on 2 NVIDIA A100 GPUs with 40GB memory each. We employed a SigLIP [48] visual encoder and a LLaMA-2-7B language model [13]. The LLM layers utilize low-rank adaptation [16] with rank $r = 16$ to minimize parameter overhead, and avoid full fine-tuning of the LLM. Training was performed using the AdamW optimizer with a learning rate of 2×10^{-5} and a cosine annealing schedule [24] over 50,000 steps. Gradient accumulation over 16 steps was employed to effectively increase batch size while managing memory constraints.

4 Experiments and Results

This section presents and discusses the experimental evaluations performed to measure unimodal bias across state-of-the-art multimodal Vision Language Models (VLMs). We evaluated several leading models under various scenarios using the MMWorld dataset, as detailed in Chapter 3, along with additional experiments using image grids to facilitate comparisons across different architectures.

4.1 Experimental Setup

The MMWorld dataset [15] allowed us to systematically assess modality bias using five distinct scenarios:

1. Matching Captions & Matching Videos
2. Mismatched Captions & Matching Videos
3. Blank Video & Matching Captions
4. No Caption & Matching Video
5. Matching Captions & Mismatched Videos

For image-language models, we constructed image grids to simulate video input, which is demonstrated to be an effective proxy for video data, often yielding superior performance compared to traditional video language models in zero-shot tasks [19].

4.2 Experimental Results

Tables 1a and 1b present performance results of various state-of-the-art video language models evaluated on the MMWorld dataset using non-prompted and prompted setups. The models are assessed using three metrics: Text Bias Index (TBI), Visual Bias Index (VBI), and Modality Reliance Score (MRS).

The non-prompted evaluation results in Table 1a reveal a strong reliance on textual inputs across most models. Text Bias Index (TBI) scores are substantially higher than Visual Bias Index (VBI), suggesting a prevalent unimodal bias. Notably, models like Video LLaVA (TBI = 0.520) and VoRA (TBI = 0.659) show high textual dependence, with correspondingly high MRS values, indicating difficulty in leveraging visual information effectively. In contrast, Qwen-7b and Gemma3-12b display more balanced behavior with lower MRS values of 9.213 and 10.012, respectively.

Prompted evaluations shown in Table 1b introduce strategies such as text prefixes and cross-attention, which lead to moderate improvements in modality balance. TBI values decrease slightly while VBI values generally stay around the same range, resulting

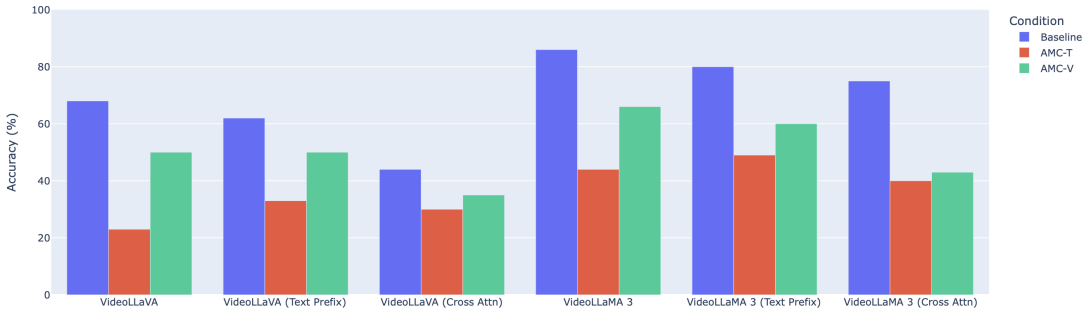


Figure 3: Model accuracy across different input conditions (baseline, mismatched text (AMC-T), and mismatched videos (AMC-V)) showing the impact of modality disruptions and the efficacy of prompting techniques (e.g., text prefixing and cross-attention) in improving robustness.

in reduced MRS scores across most models. These changes indicate improved reliance on visual cues and a more balanced integration of modalities. For instance, VideoLLaMA 3 shows a reduction in MRS from 23.103 to 16.895 after prompting, suggesting that prompting strategies can help mitigate unimodal bias, but not totally eliminating them.

Model	TBI	VBI	MRS
VideoLLaMA 3 [49]	0.388	0.152	23.103
Chat Uni-Vi [17]	0.296	0.017	17.642
Video LLaVA [22]	0.520	0.043	26.785
VideoGPT+ [26]	0.351	0.069	19.616
Qwen2.5-VL [5]	0.432	0.088	9.213
SmolVLM2 [27]	0.408	0.102	15.878
LLaMA 3.2 [13]	0.461	0.112	16.259
Gemma3-12b [38]	0.426	0.097	10.012
VoRA [45]	0.659	0.074	55.241

(a) Non-prompted evaluation metrics showing dominant text modality reliance.

Model	TBI	VBI	MRS
VideoLLaMA 3 [49]	0.365	0.176	16.895
Chat Uni-Vi [17]	0.278	0.059	13.991
Video LLaVA [22]	0.493	0.108	20.002
VideoGPT+ [26]	0.328	0.123	15.219
Qwen2.5-VL [5]	0.398	0.122	7.764
SmolVLM2 [27]	0.372	0.136	12.386
LLaMA 3.2 [13]	0.417	0.138	14.931
Gemma3-12b [38]	0.399	0.142	8.745
VoRA [45]	0.604	0.128	53.951

(b) Prompted evaluation metrics showing modest improvement in modality balance.

4.3 Analysis and Discussion

The experimental results clearly demonstrate that current multimodal models exhibit significant unimodal bias, especially favoring textual modalities. In non-prompted conditions, models show a pronounced drop in performance when provided with mismatched captions, underscoring their overreliance on textual cues for decision-making.

Comparing our results with our initial hypotheses, we find several important insights. As predicted, prompting strategies produced modest improvements in modality balance across models, but did not fully resolve the underlying bias issues. The Text Bias Index (TBI) decreased only slightly with prompting (e.g., from 0.388 to 0.365 for VideoLLaMA 3), confirming our expectation that deeply embedded biases cannot be fully

addressed through instruction alone.

The Interleaved Vision-Text Projection (IVTP) approach yielded more substantial improvements in Visual Bias Index (VBI) compared to simple prompting, supporting our hypothesis that architectural modifications addressing token arrangement would enhance modality integration. However, we also observed the trade-off we anticipated, with a slight decrease in overall baseline accuracy, suggesting that forcing balanced modality processing may come at a cost to performance in some scenarios.

The Cross-Attention Projector demonstrated the most significant reduction in Modality Reliance Score (MRS) among our approaches, confirming our hypothesis that enabling dynamic, content-dependent interactions between modalities would provide the most effective bias mitigation. However, this approach also required the most computational resources and fine-tuning iterations, aligning with our expectation regarding implementation complexity.

Interestingly, while all bias mitigation strategies improved modality balance to some degree, they also resulted in slight decreases in overall baseline accuracy. This suggests that there may be an inherent trade-off between optimizing for balanced modality integration and maximizing task-specific performance. This trade-off is particularly evident in the reductions in Adversarial Modality Conflict (AMC-V and AMC-T) accuracies across most models in the adjusted setups. For instance, in fig. 3, we can see for VideoLLaMA 3 the difference between AMC-V and AMC-T narrows with our mitigation approaches, but at a small cost to baseline performance.

Additionally, the use of image grids as video input proxies proves effective in enhancing performance for image-based models [19], often outperforming their video-based counterparts in zero-shot tasks. This highlights the potential of image grid methodologies as a more efficient and sometimes superior alternative to conventional video input processing.

In summary, these experiments not only confirm the presence of strong modality biases in current VLMs but also validate our hypothesized effectiveness hierarchy of mitigation strategies: Cross-Attention Projector Interleaved Vision-Text Projection Prompting. These findings reinforce the need for deliberate architectural modifications to foster truly balanced multimodal understanding, rather than relying solely on prompt engineering approaches.


VideoLLaVA (Baseline)
VideoLLaVA (AMC-T)
VideoLLaVA (AMC-V)
VideoLLaVA (Text Prefix) (Baseline)
VideoLLaVA (Text Prefix) (AMC-T)
VideoLLaVA (Text Prefix) (AMC-V)
VideoLLaVA (Cross Attn) (Baseline)
VideoLLaVA (Cross Attn) (AMC-T)
VideoLLaVA (Cross Attn) (AMC-V)
VideoLLaMA 3 (Baseline)
VideoLLaMA 3 (AMC-T)
VideoLLaMA 3 (AMC-V)
VideoLLaMA 3 (Text Prefix) (Baseline)
VideoLLaMA 3 (Text Prefix) (AMC-T)
VideoLLaMA 3 (Text Prefix) (AMC-V)
VideoLLaMA 3 (Cross Attn) (Baseline)
VideoLLaMA 3 (Cross Attn) (AMC-T)
VideoLLaMA 3 (Cross Attn) (AMC-V)
Tech & Engineering
Sports & Arts
Science
Health & Medicine
Games
Embodied Tasks
Business
Count
25
20
15
10
5

Figure 4

5 Conclusion

5.1 Summary of Contributions

In this thesis, we systematically explored unimodal bias within Vision Language Models (VLMs), confirming a prevalent overreliance on textual information. Our analyses using the MMWorld dataset revealed significant performance degradation when models encountered conflicting modality inputs or absent textual cues, emphasizing the practical implications of modality imbalance. By employing targeted bias mitigation techniques—including explicit Prompting, Interleaved Vision-Text Projection (IVTP), and Cross-Attention Projector architectures—we achieved demonstrable improvements in multimodal balance and overall robustness. These approaches provide a promising framework for addressing modality bias, ultimately enhancing the performance and fairness of multimodal models.

5.2 Implications and Limitations

The findings indicate that while textual modality often provides explicit, easily accessible semantic information, models must be actively guided and architecturally designed to appropriately leverage visual modalities, particularly in ambiguous or conflicting contexts. Our proposed methodologies offer significant advancements toward balanced multimodal reasoning, though further exploration is necessary to fully realize their potential in broader real-world scenarios. One notable limitation of this study is its reliance on existing datasets, as there are relatively few specifically designed to evaluate modality biases. Moreover, our experiments were constrained by utilizing only a subset of approximately 400 videos from the MMWorld dataset, limiting the scope and generalizability of our findings. Future work would greatly benefit from developing and employing more extensive datasets that better capture the complexity of multimodal interactions.

5.3 Future Research Directions

Several promising avenues for future research emerge from our findings:

5.3.1 GRPO with Vision Language Models

Future research could leverage Gradient-based Representation Projection Optimization (GRPO) [35] in conjunction with VLMs to enhance multimodal reasoning capabilities. This approach involves explicitly guiding the gradient flow between modalities during training, potentially decreasing modality biases by aligning representation spaces more effectively. Integrating GRPO could enable models to better reason with nuanced interactions between visual and textual cues, significantly reducing unimodal dominance. Data for this setup could be generated using a external larger model API which can create the

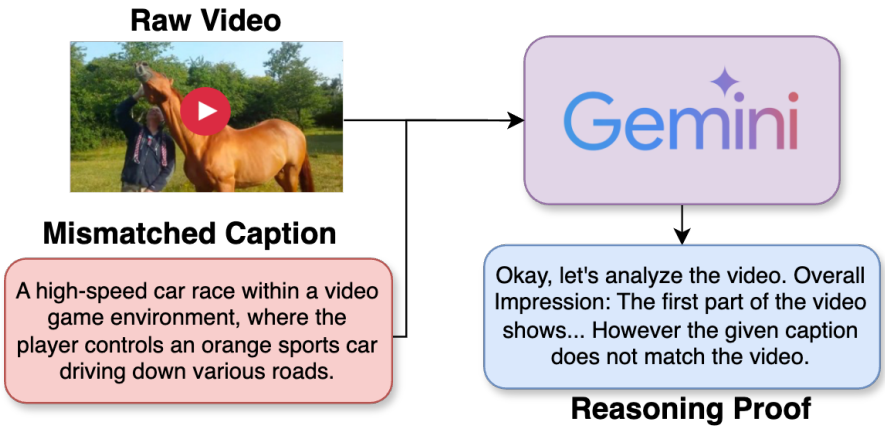


Figure 5: Data Generation Setup for GRPO.

training dataset for this GRPO setup as shown in 5.

5.3.2 Dataset Development for Unimodal Bias Evaluation

The creation of dedicated datasets specifically designed to evaluate unimodal bias remains underexplored. Developing such datasets, which explicitly challenge models to simultaneously interpret and integrate conflicting or complementary visual and textual inputs, could greatly enhance the effectiveness and generalizability of bias mitigation techniques. These datasets would enable deeper investigation into the nuanced dynamics of multimodal integration.

5.3.3 Optimal Transport for Multimodal Alignment

Optimal transport methods offer a compelling mathematical framework to address distributional mismatches between visual and textual modalities. Integrating optimal transport into the loss functions or training strategies of VLMs could promote better alignment between modality distributions, potentially enhancing both the interpretability and the effectiveness of multimodal fusion processes. Exploring optimal transport methodologies specifically tailored for VLMs could yield significant improvements in multimodal coherence and bias reduction.

Theoretical foundation. Let μ_V and μ_T denote latent distributions of visual and textual tokens, respectively. Minimising $W_2(\mu_V, \mu_T)$ is equivalent to finding a coupling $\pi^\star$ that describes how probability mass should "flow" from vision to language space with minimal cost. The Sinkhorn-regularised solution

$$\pi^\star = \arg\min_{\pi \in \Pi(\mu_V, \mu_T)} \langle \pi,\, C \rangle + \varepsilon H(\pi),$$

where $C_{ij} = ||v_i - t_j||_2^2$, is differentiable, enabling end-to-end optimisation inside large transformers. Importantly, the $\varepsilon H(\pi)$ term ensures numerical stability and linear-time GPU kernels.

Practical recipe for VLMs.

- **Step 1 (Pre-alignment)**: Collect an auxiliary corpus in which each image–text pair is weakly matched (e.g., COCO captions [23]). Train f_θ to minimise OTDD between visual features and text embeddings of *that same pair*; this initialises a geometry where modalities overlap.
- **Step 2 (Task-specific fine-tuning)**: On the target dataset, jointly minimise

$$\mathcal{L}_{total} = \mathcal{L}_{\text{CE}} + \beta \mathcal{L}_{\text{OT}} + \gamma \mathcal{L}_{\text{reg}}.$$

- **Step 3 (Post-hoc calibration).**: For classification tasks with ambiguous cues, compute an OT barycentre of softmax outputs from uni-modal ablations; this acts as a modality-aware ensemble that preserves minority-modal evidence.

Expected benefits. OT alignment is bias-agnostic: it does not depend on synthetic counterfactuals and thus scales to open-vocabulary VLMs.

Future experiments. We propose three concrete lines of inquiry:

- **Entropy-conditioned transport**: Dynamically adjust the regularisation ε based on per-batch modality entropy, encouraging sharper couplings when vision provides decisive evidence.
- **Temporal OT for video-language**: Define a cost on frame positions to enforce causal matching, extending work on sequence OT in speech recognition.

Risks & limitations: OT hinges on a meaningful ground metric; in transformer spaces, Euclidean distance may not reflect semantic proximity. Recent progress in learned ground metrics—where C is itself an MLP with Lipschitz constraint—could alleviate this, but theoretical guarantees remain open. Computationally, although Sinkhorn scales quadratically, extreme-resolution video still challenges GPU memory; stochastic minibatch OT and divergence-free vector fields are promising mitigations.

These future directions represent critical next steps towards achieving balanced, equitable, and robust multimodal models, addressing inherent biases and unlocking new possibilities in multimodal AI.

References

[1] Aishwarya Agrawal et al. *Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering*. 2018. arXiv: 1712.00377 [cs.CV]. URL: https://arxiv.org/abs/1712.00377.

[2] Aishwarya Agrawal et al. *VQA: Visual Question Answering*. 2016. arXiv: 1505.00468 [cs.CL]. URL: https://arxiv.org/abs/1505.00468.

[3] Jean-Baptiste Alayrac et al. *Flamingo: a Visual Language Model for Few-Shot Learning*. 2022. arXiv: 2204.14198 [cs.CV]. URL: https://arxiv.org/abs/2204.14198.

[4] David Alvarez-Melis and Nicolò Fusi. *Geometric Dataset Distances via Optimal Transport*. 2020. arXiv: 2002.02923 [cs.LG]. URL: https://arxiv.org/abs/2002.02923.

[5] Shuai Bai et al. *Qwen2.5-VL Technical Report*. 2025. arXiv: 2502.13923 [cs.CV]. URL: https://arxiv.org/abs/2502.13923.

[6] Remi Cadene et al. *RUBi: Reducing Unimodal Biases in Visual Question Answering*. 2020. arXiv: 1906.10169 [cs.CV]. URL: https://arxiv.org/abs/1906.10169.

[7] Meiqi Chen et al. *Quantifying and Mitigating Unimodal Biases in Multimodal Large Language Models: A Causal Perspective*. 2024. arXiv: 2403.18346 [cs.CL]. URL: https://arxiv.org/abs/2403.18346.

[8] Meiqi Chen et al. *Quantifying and Mitigating Unimodal Biases in Multimodal Large Language Models: A Causal Perspective*. 2024. arXiv: 2403.18346 [cs.CL]. URL: https://arxiv.org/abs/2403.18346.

[9] Tsai-Shien Chen et al. *Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers*. 2024. arXiv: 2402.19479 [cs.CV]. URL: https://arxiv.org/abs/2402.19479.

[10] Gabriele Dominici et al. *SHARCS: Shared Concept Space for Explainable Multimodal Learning*. 2023. arXiv: 2307.00316 [cs.LG]. URL: https://arxiv.org/abs/2307.00316.

[11] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV]. URL: https://arxiv.org/abs/2010.11929.

[12] Ziv Goldfeld et al. *Limit distribution theory for smooth p-Wasserstein distances*. 2022. arXiv: 2203.00159 [math.PR]. URL: https://arxiv.org/abs/2203.00159.

[13] Aaron Grattafiori et al. *The Llama 3 Herd of Models*. 2024. arXiv: 2407.21783 [cs.AI]. URL: https://arxiv.org/abs/2407.21783.

[14] Suchin Gururangan et al. *Annotation Artifacts in Natural Language Inference Data*. 2018. arXiv: 1803.02324 [cs.CL]. URL: https://arxiv.org/abs/1803.02324.

[15] Xuehai He et al. *MMWorld: Towards Multi-discipline Multi-faceted World Model Evaluation in Videos*. 2024. arXiv: 2406.08407 [cs.CV]. URL: https://arxiv.org/abs/2406.08407.

[16] Edward J. Hu et al. *LoRA: Low-Rank Adaptation of Large Language Models*. 2021. arXiv: 2106.09685 [cs.CL]. URL: https://arxiv.org/abs/2106.09685.

[17] Peng Jin et al. *Chat-UniVi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding*. 2024. arXiv: 2311.08046 [cs.CV]. URL: https://arxiv.org/abs/2311.08046.

[18] Douwe Kiela et al. *The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes*. 2021. arXiv: 2005.04790 [cs.AI]. URL: https://arxiv.org/abs/2005.04790.

[19] Wonkyun Kim et al. *An Image Grid Can Be Worth a Video: Zero-shot Video Question Answering Using a VLM*. 2024. arXiv: 2403.18406 [cs.CV]. URL: https://arxiv.org/abs/2403.18406.

[20] JuneHyoung Kwon et al. *See-Saw Modality Balance: See Gradient, and Sew Impaired Vision-Language Balance to Mitigate Dominant Modality Bias*. 2025. arXiv: 2503.13834 [cs.CV]. URL: https://arxiv.org/abs/2503.13834.

[21] Zichao Li et al. *The Devil Is in the Details: Tackling Unimodal Spurious Correlations for Generalizable Multimodal Reward Models*. 2025. arXiv: 2503.03122 [cs.CL]. URL: https://arxiv.org/abs/2503.03122.

[22] Bin Lin et al. *Video-LLaVA: Learning United Visual Representation by Alignment Before Projection*. 2024. arXiv: 2311.10122 [cs.CV]. URL: https://arxiv.org/abs/2311.10122.

[23] Tsung-Yi Lin et al. *Microsoft COCO: Common Objects in Context*. 2015. arXiv: 1405.0312 [cs.CV]. URL: https://arxiv.org/abs/1405.0312.

[24] Ilya Loshchilov and Frank Hutter. *SGDR: Stochastic Gradient Descent with Warm Restarts*. 2017. arXiv: 1608.03983 [cs.LG]. URL: https://arxiv.org/abs/1608.03983.

[25] Jiasen Lu et al. *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*. 2019. arXiv: 1908.02265 [cs.CV]. URL: https://arxiv.org/abs/1908.02265.

[26] Muhammad Maaz et al. *VideoGPT+: Integrating Image and Video Encoders for Enhanced Video Understanding*. 2024. arXiv: 2406.09418 [cs.CV]. URL: https://arxiv.org/abs/2406.09418.

[27] Andrés Marafioti et al. *SmolVLM: Redefining small and efficient multimodal models*. 2025. arXiv: 2504.05299 [cs.AI]. URL: https://arxiv.org/abs/2504.05299.

[28] Eduardo Fernandes Montesuma, Fred Ngolè Mboula, and Antoine Souloumiac. *Recent Advances in Optimal Transport for Machine Learning*. 2024. arXiv: 2306.16156 [cs.LG]. URL: https://arxiv.org/abs/2306.16156.

[29] Yongyu Mu et al. *Cross-layer Attention Sharing for Large Language Models*. 2024. arXiv: 2408.01890 [cs.CL]. URL: https://arxiv.org/abs/2408.01890.

[30] Arsha Nagrani et al. *Attention Bottlenecks for Multimodal Fusion*. 2022. arXiv: 2107.00135 [cs.CV]. URL: https://arxiv.org/abs/2107.00135.

[31] OpenAI et al. *GPT-4 Technical Report*. 2024. arXiv: 2303.08774 [cs.CL]. URL: https://arxiv.org/abs/2303.08774.

[32] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV]. URL: https://arxiv.org/abs/2103.00020.

[33] Mrigank Rochan and Yang Wang. *Video Summarization by Learning from Unpaired Data*. 2019. arXiv: 1805.12174 [cs.CV]. URL: https://arxiv.org/abs/1805.12174.

[34] Pranab Sahoo et al. *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications*. 2025. arXiv: 2402.07927 [cs.AI]. URL: https://arxiv.org/abs/2402.07927.

[35] Zhihong Shao et al. *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models*. 2024. arXiv: 2402.03300 [cs.CL]. URL: https://arxiv.org/abs/2402.03300.

[36] Junhong Shen et al. *Cross-Modal Fine-Tuning: Align then Refine*. 2023. arXiv: 2302.05738 [cs.LG]. URL: https://arxiv.org/abs/2302.05738.

[37] Zhongtian Sun et al. "Is Unimodal Bias Always Bad for Visual Question Answering? A Medical Domain Study with Dynamic Attention". In: Dec. 2022, pp. 5352–5360. DOI: 10.1109/BigData55660.2022.10020791.

[38] Gemma Team et al. *Gemma 3 Technical Report*. 2025. arXiv: 2503.19786 [cs.CL]. URL: https://arxiv.org/abs/2503.19786.

[39] Gemma Team et al. *Gemma: Open Models Based on Gemini Research and Technology*. 2024. arXiv: 2403.08295 [cs.CL]. URL: https://arxiv.org/abs/2403.08295.

[40] Changyao Tian et al. *MM-Interleaved: Interleaved Image-Text Generative Modeling via Multi-modal Feature Synchronizer*. 2024. arXiv: 2401.10208 [cs.CV]. URL: https://arxiv.org/abs/2401.10208.

[41] Hugo Touvron et al. *LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: 2302.13971 [cs.CL]. URL: https://arxiv.org/abs/2302.13971.

[42] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: https://arxiv.org/abs/1706.03762.

[43] Oriol Vinyals et al. *Show and Tell: A Neural Image Caption Generator*. 2015. arXiv: 1411.4555 [cs.CV]. URL: https://arxiv.org/abs/1411.4555.

[44] Shakti N. Wadekar et al. *The Evolution of Multimodal Model Architectures*. 2024. arXiv: 2405.17927 [cs.AI]. URL: https://arxiv.org/abs/2405.17927.

[45] Han Wang et al. *Vision as LoRA*. 2025. arXiv: 2503.20680 [cs.CV]. URL: https://arxiv.org/abs/2503.20680.

[46] Renjie Wu et al. *Deep Multimodal Learning with Missing Modality: A Survey*. 2024. arXiv: 2409.07825 [cs.CV]. URL: https://arxiv.org/abs/2409.07825.

[47] Lingling Xu et al. *Parameter-Efficient Fine-Tuning Methods for Pretrained Language Models: A Critical Review and Assessment*. 2023. arXiv: 2312.12148 [cs.CL]. URL: https://arxiv.org/abs/2312.12148.

[48] Xiaohua Zhai et al. *Sigmoid Loss for Language Image Pre-Training*. 2023. arXiv: 2303.15343 [cs.CV]. URL: https://arxiv.org/abs/2303.15343.

[49] Boqiang Zhang et al. *VideoLLaMA 3: Frontier Multimodal Foundation Models for Image and Video Understanding*. 2025. arXiv: 2501.13106 [cs.CV]. URL: https://arxiv.org/abs/2501.13106.

[50] Yedi Zhang, Peter E. Latham, and Andrew Saxe. *Understanding Unimodal Bias in Multimodal Deep Linear Networks*. 2024. arXiv: 2312.00935 [cs.LG]. URL: https://arxiv.org/abs/2312.00935.