

Beyond the Efficiency Paradox: Unlocking Multimodal Subspaces via Cumulative Information Span

Xiuze Zhou¹, Xiaowen Chu¹, and Xuming Hu^{1*}

¹The Hong Kong University of Science and Technology (Guangzhou)

xz.zhou@connect.hkust-gz.edu.cn, {xwchu, xuminghu}@hkust-gz.edu.cn

Abstract

In real-world multimodal applications, different modalities often exhibit weakly aligned adaptation requirements, making it challenging for a shared adapter to effectively capture heterogeneous signals. Existing approaches typically use a unified low-rank adapter for concatenated multimodal representations, leading to competition among modality-specific adaptations and diminishing returns as the adapter rank increases. To address this limitation, we propose msLoRA, a modality-specific adaptation framework for efficient multimodal fine-tuning. msLoRA assigns independent low-rank adapters to different modalities, allowing them to capture specialized adaptation directions while maintaining efficient parameterization. Our theoretical analysis introduces the concept of cumulative information span, demonstrating that modality-specific adapters can expand the effective adaptation space and avoid the cross-modal disagreement typical of shared adapters. We evaluate msLoRA using multiple prominent open-source LLMs, including the Llama, Qwen, and Mistral families, across four multimodal tasks, showing consistent improvements in scalability and effectiveness for low-rank adaptation in multimodal LLMs. The source code is available at <https://github.com/XiuzeZhou/msLoRA/>.

1 Introduction

The rapid evolution of Large Language Models (LLMs), exemplified by breakthroughs such as ChatGPT and GPT-4, has profoundly transformed the artificial intelligence landscape (Chen et al., 2025). These models have driven substantial advances across various domains (Lifan et al., 2026; Chang et al., 2024; Ding et al., 2023; Chen et al., 2022). With their ability to process, understand,

and generate human-like text, LLMs have demonstrated remarkable versatility in applications ranging from conversational agents to content creation (Wölflein et al., 2025; Wu et al., 2024).

For example, in video content understanding, LLMs have emerged as powerful tools by leveraging their capacity for multimodal reasoning and temporal alignment (Yuan et al., 2025; Shu et al., 2025; Huang et al., 2024). Similarly, in multimodal sentiment analysis, LLMs bridge heterogeneous modalities to assess sentiment polarity and capture nuanced human perspectives (Yang et al., 2025a; Mu et al., 2024; Li et al., 2025c). Despite their promise, existing LLM-based multimodal methods still face two major challenges:

Suboptimal Multimodal Integration. In many real-world applications, the relationships between multimodal inputs are complex and weakly correlated (Zhang et al., 2023; Liu et al., 2024). Such weak cross-modal correlation poses challenges for conventional approaches, which often struggle to integrate diverse signals while preserving modality-specific information. Mathematically, forcing heterogeneous updates into a single shared low-rank subspace can cause interference among modality-specific features, impairing the model’s ability to represent distinct data types.

Prohibitive Computational Costs and the Efficiency Paradox. The massive scale of modern LLMs makes full-parameter fine-tuning impractical in resource-constrained environments (Kabongo et al., 2025; Li et al., 2025b,d). Even with Parameter-Efficient Fine-Tuning (PEFT) techniques like LoRA, performance does not always improve with increasing adapter ranks, a phenomenon we call the *efficiency paradox* (Jung et al., 2025). Beyond a suitable rank, additional dimensions may capture redundant or noisy directions, leading to diminishing returns or even empirical performance degradation.

*Corresponding author.

To address these challenges, we propose msLoRA, which mitigates the efficiency paradox through decentralized adaptation. Instead of a single shared adapter, msLoRA assigns separate low-rank adapters to heterogeneous data streams. By updating a small set of modality-specific low-rank matrices, msLoRA captures complementary information from different modalities while reducing interference in a shared adaptation space.

In our theoretical analysis, we introduce the concept of cumulative information span. Under approximate modality orthogonality, modality-specific updates can preserve complementary spectral components, allowing adaptation directions to accumulate rather than compete within a single shared subspace. In the exact orthogonal case, the effective rank of msLoRA reaches the cumulative sum of modality-specific ranks, providing a broader update span than a shared LoRA adapter. Our analysis further shows that decentralized rank allocation can achieve lower approximation error within the same cumulative rank budget.

Empirically, we conducted extensive experiments across four distinct multimodal tasks using a range of open-source LLMs, including the Llama, Qwen, and Mistral families. The results show that msLoRA achieves strong and stable performance across different backbones and tasks, often matching or exceeding unified LoRA and other PEFT baselines with lower per-adapter nominal ranks. These findings support our modality-specific adaptation hypothesis and suggest that decentralized rank allocation is effective for multimodal integration. Overall, msLoRA offers a scalable and robust solution for real-world applications that require a balance between computational efficiency and multimodal representation quality.

Our contributions are summarized as follows:

- We propose msLoRA, a modality-specific low-rank adaptation framework that integrates heterogeneous multimodal signals within LLMs while reducing inter-modality interference.
- We introduce a decentralized rank allocation strategy that improves parameter efficiency and mitigates the efficiency paradox observed at large adapter ranks.
- We provide a theoretical analysis based on cumulative information span and support it with SVD spectral analysis and gradient similarity measurements.

2 Related Work

2.1 Multimodal Large Language Models

The integration of multimodal features into LLMs primarily focuses on aligning heterogeneous data types—such as images, audio, and video—into a unified semantic space. Predominant architectures utilize learnable projection layers to map modality-specific embeddings into the LLM’s input domain. In video understanding, frameworks such as VideoRefer (Yuan et al., 2025), Audio-Visual LLM (Shu et al., 2025), and VTimeLLM (Huang et al., 2024) fuse disparate data streams to capture complex temporal dynamics. Despite their advancements, these methods face two limitations:

- **Computational Inefficiency:** The large volume of multimodal data, combined with the increasing scale of modern LLM weights, makes full-parameter fine-tuning impractical in many resource-constrained settings (Kabongo et al., 2025; Li et al., 2025a,d).
- **Modality Interference:** Existing models fail to account for the inherent heterogeneity of modalities. Forcing diverse features into a single, shared subspace leads to feature collisions and gradient interference, compromising training stability and representation quality (Yamanishi et al., 2024; Radford et al., 2021).

2.2 Parameter-Efficient Fine-Tuning (PEFT)

To reduce the costs of full-parameter updates, PEFT strategies like LoRA (Hu et al., 2021) use low-rank decomposition for weight updates. However, recent studies reveal an *efficiency paradox*: performance does not consistently improve with higher ranks (Yang et al., 2025b; Jung et al., 2025). Exceeding an optimal threshold leads to diminishing returns due to subspace overfitting (Wang et al., 2025) and gradient redundancy (Wang et al., 2024).

Although some LoRA-based methods, such as MTL-LoRA (Yang et al., 2025b) and GraLoRA (Jung et al., 2025), improve rank allocation, they are primarily designed for unified adaptation without explicitly separating modality-specific update directions. As a result, heterogeneous multimodal gradients may still compete within a shared low-rank subspace, leading to performance degradation. In contrast, msLoRA introduces the concept of *cumulative information span*, where decentralized adapters preserve complementary modality-specific components.

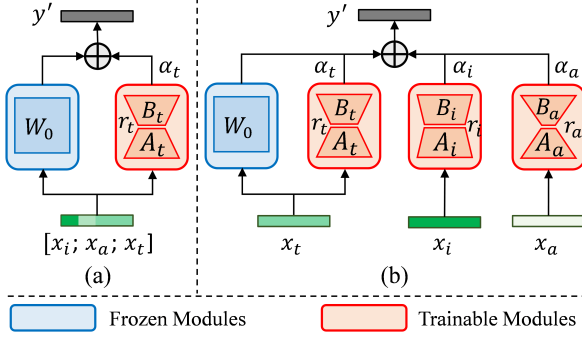


Figure 1: Comparison between (a) standard LoRA and (b) our proposed msLoRA. For multimodal inputs (image i , audio a , and text t), standard LoRA applies a shared low-rank adaptation operator to concatenated multimodal representations $[x_i; x_a; x_t]$, while msLoRA uses modality-specific adapters with ranks r_m and scaling factors α_m to capture distinct adaptation directions.

3 Proposed Method: msLoRA

To address the efficiency paradox observed in standard LoRA fine-tuning, we propose msLoRA, a modality-specific adaptation framework for multimodal tasks. Instead of forcing heterogeneous modalities to share a single low-rank adaptation operator, msLoRA assigns decentralized LoRA modules to different modalities, enabling modality-specific adaptation while reducing inter-modality interference. The architecture is shown in Figure 1.

For each modality m , msLoRA introduces low-rank factors $A_m \in \mathbb{R}^{r_m \times d}$ and $B_m \in \mathbb{R}^{d \times r_m}$. During fine-tuning, only these low-rank matrices and modality projection layers are optimized. The update for each modality is defined as follows:

$$\Delta W_m = \alpha_m B_m A_m, \quad (1)$$

where α_m denotes the effective scaling factor of modality m .

To integrate the representations learned by different modality-specific modules, their adapter outputs are fused into a shared latent state for downstream generative tasks. For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times d}$, the update of msLoRA is defined as follows:

$$y' = (W_0 + \Delta W_t)x_t + \sum_{m \in \mathcal{M} | m \neq t} \Delta W_m x_m, \quad (2)$$

where x_t denotes a textual representation; $\mathcal{M}$ denotes the set of modalities; and x_m denotes the corresponding modality representation. For each non-text modality, $x_m = \text{Norm}(f_m(e_m))$ denotes its projected and normalized representation in the LLM latent space.

4 Theoretical Analysis

We analyze msLoRA from the perspective of low-rank approximation. Let $\mathcal{M}$ denote the set of adapted modalities with $|\mathcal{M}| = M$. For modality $m \in \mathcal{M}$, let $G_m \in \mathbb{R}^{d \times d}$ denote an ideal modality-specific target update. A unified LoRA adapter uses one shared low-rank update to approximate all $\{G_m\}_{m=1}^M$, whereas msLoRA assigns an independent low-rank update ΔW_m to each modality. Since W_0 is frozen and shared, our analysis focuses only on the modality-dependent residual updates.

4.1 Efficiency Paradox

Increasing the LoRA rank does not always improve empirical performance (Hu et al., 2021; Yang et al., 2025b; Jung et al., 2025). Beyond a certain rank, extra dimensions may capture noisy or redundant directions, leading to diminishing returns. This is exacerbated in multimodal settings, where heterogeneous modalities may induce weakly correlated adaptation directions. A shared adapter must approximate these updates using the same adaptation operator, leading to competition among modality-specific adaptation directions.

Our analysis does not imply that approximation error increases with rank. Instead, it shows that increasing the rank of a shared adapter can reduce truncation error but cannot eliminate the error induced by cross-modal disagreement. Thus, the analysis explains rank saturation rather than the empirical degradation seen at excessively large ranks.

4.2 Cumulative Information Span

Let the compact SVD of the modality-specific target update be

$$G_m = U_m \Sigma_m V_m^\top, \quad (3)$$

$$\Sigma_m = \text{diag}(\sigma_{m,1}, \dots, \sigma_{m,s_m}),$$

where $\sigma_{m,1} \geq \dots \geq \sigma_{m,s_m} > 0$ and $s_m = \text{rank}(G_m)$. Denote the best rank- r_m approximation of G_m by $G_m^{(r_m)}$, where $0 \leq r_m \leq s_m$.

Because msLoRA applies different updates to different modality inputs, we analyze the joint multimodal operator

$$\mathcal{G} = [G_1, \dots, G_M] \in \mathbb{R}^{d \times Md}, \quad (4)$$

$$z = [x_1^\top, \dots, x_M^\top]^\top,$$

so that

$$\sum_{m=1}^M G_m x_m = \mathcal{G} z. \quad (5)$$

The corresponding idealized msLoRA approximation is

$$\mathcal{G}_{\text{ms}} = [G_1^{(r_1)}, \dots, G_M^{(r_M)}]. \quad (6)$$

Assumption 1. For distinct modalities $m \neq n$, the retained dominant output singular subspaces are approximately orthogonal:

$$U_{m,r_m}^\top U_{n,r_n} \approx 0. \quad (7)$$

The exact case is used in Lemma 1, while experiments in Section 5.3 provide empirical evidence related to this assumption.

Lemma 1. Under the exact version of Assumption 1, if $K = \sum_{m=1}^M r_m \leq d$, then

$$\text{rank}(\mathcal{G}_{\text{ms}}) = K. \quad (8)$$

Moreover, the nonzero singular values of $\mathcal{G}_{\text{ms}}$ are exactly the multiset union of the retained singular values of all modality-specific approximations.

Proof. For each modality,

$$G_m^{(r_m)} = U_{m,r_m} \Sigma_{m,r_m} V_{m,r_m}^\top. \quad (9)$$

Let

$$\begin{aligned} U &= [U_{1,r_1}, \dots, U_{M,r_M}], \\ \bar{V} &= \text{blkdiag}(V_{1,r_1}, \dots, V_{M,r_M}). \end{aligned} \quad (10)$$

By Assumption 1, U has orthonormal columns. Since different modalities occupy disjoint input blocks, $\bar{V}$ also has orthonormal columns. Therefore, we can write

$$\mathcal{G}_{\text{ms}} = U \text{blkdiag}(\Sigma_{1,r_1}, \dots, \Sigma_{M,r_M}) \bar{V}^\top. \quad (11)$$

This is an SVD up to a permutation of singular values. Thus, $\mathcal{G}_{\text{ms}}$ has exactly K nonzero singular values, and hence $\text{rank}(\mathcal{G}_{\text{ms}}) = K$.

4.3 Shared-Adapter Interference and Rank Allocation

A unified rank- r adapter applies the same matrix H , with $\text{rank}(H) \leq r$, to all modality inputs. Its joint operator is

$$\mathcal{H}_{\text{uni}} = [H, \dots, H], \quad (12)$$

and therefore

$$\|\mathcal{G} - \mathcal{H}_{\text{uni}}\|_F^2 = \sum_{m=1}^M \|G_m - H\|_F^2 \triangleq \mathcal{E}_{\text{uni}}(H). \quad (13)$$

Let $\bar{G} = \frac{1}{M} \sum_{m=1}^M G_m$. Then the unified objective decomposes as

$$\mathcal{E}_{\text{uni}}(H) = \sum_{m=1}^M \|G_m - \bar{G}\|_F^2 + M \|\bar{G} - H\|_F^2. \quad (14)$$

By the Eckart-Young-Mirsky theorem (Golub et al., 1987), the optimal rank- r unified adapter satisfies

$$\mathcal{E}_{\text{uni}}^*(r) = \underbrace{\sum_{m=1}^M \|G_m - \bar{G}\|_F^2}_{\mathcal{D}_{\text{modal}}} + M \sum_{i=r+1}^d \bar{\sigma}_i^2, \quad (15)$$

where $\bar{\sigma}_i$ are the singular values of $\bar{G}$. The first term $\mathcal{D}_{\text{modal}}$ measures cross-modal disagreement and is independent of r . Thus, increasing the rank of a shared adapter can reduce the truncation error but cannot remove this disagreement term.

When the stacked latent input is jointly whitened, i.e., $\mathbb{E}[zz^\top] = I$, the Frobenius objective also equals the expected squared output error. In contrast, the optimal branch-wise approximation error of msLoRA is

$$\begin{aligned} \mathcal{E}_{\text{ms}}(\{r_m\}) &= \sum_{m=1}^M \|G_m - G_m^{(r_m)}\|_F^2 \\ &= \sum_{m=1}^M \sum_{i=r_m+1}^{s_m} \sigma_{m,i}^2. \end{aligned} \quad (16)$$

For a fixed cumulative rank budget $K = \sum_m r_m$, increasing $r_m < s_m$ by one reduces the error by

$$\Delta \mathcal{E}_m = \sigma_{m,r_m+1}^2. \quad (17)$$

Hence, since each singular spectrum is nonincreasing, the optimal allocation greedily selects the largest remaining spectral component:

$$m^* = \arg \max_{m:r_m < s_m} \sigma_{m,r_m+1}^2. \quad (18)$$

Repeated allocation selects the globally largest K spectral components and thus minimizes the branch-wise truncation error.

Proposition 1 (Same-budget error characterization). Let $1 \leq K \leq \min(d, \sum_m s_m)$, and let $\tau_1 \geq \tau_2 \geq \dots$ denote all modality-specific singular values $\{\sigma_{m,i}\}_{m,i}$ arranged in descending order. Under the same cumulative rank budget K , the optimal errors satisfy

$$\mathcal{E}_{\text{uni}}^*(K) - \mathcal{E}_{\text{ms}}^*(K) = \sum_{j=1}^K \tau_j^2 - M \sum_{i=1}^K \bar{\sigma}_i^2, \quad (19)$$

where $\mathcal{E}_{\text{ms}}^*(K)$ denotes the minimum of $\mathcal{E}_{\text{ms}}(\{r_m\})$ over integer $0 \leq r_m \leq s_m$ and $\sum_m r_m = K$.

Consequently, under the same cumulative rank budget, msLoRA achieves no larger approximation error if and only if

$$\sum_{j=1}^K \tau_j^2 \geq M \sum_{i=1}^K \bar{\sigma}_i^2. \quad (20)$$

Proof. Optimal rank allocation gives

$$\mathcal{E}_{\text{ms}}^*(K) = \sum_{m=1}^M \|G_m\|_F^2 - \sum_{j=1}^K \tau_j^2. \quad (21)$$

Moreover, since

$$\mathcal{D}_{\text{modal}} = \sum_{m=1}^M \|G_m\|_F^2 - M \|\bar{G}\|_F^2, \quad (22)$$

the unified-adaptor error becomes

$$\mathcal{E}_{\text{uni}}^*(K) = \sum_{m=1}^M \|G_m\|_F^2 - M \sum_{i=1}^K \bar{\sigma}_i^2. \quad (23)$$

Subtracting the two expressions proves the result.

Strongly aligned targets allow a shared adapter to reuse common low-rank directions, whereas weak alignment can limit such spectral reuse.

5 Experiments

5.1 Experimental Setting

Datasets. To evaluate the generalizability and robustness of msLoRA, we use several distinct multimodal datasets spanning diverse domains:

- **ScienceQA** (Saikh et al., 2022): A multimodal **science question-answering** dataset. It evaluates multimodal commonsense reasoning and scientific knowledge understanding.
- **MVSA_Single** (Niu et al., 2016): A Twitter-based **image-text sentiment** dataset. It evaluates global multimodal sentiment classification from both textual and visual cues.
- **Twitter-17** (Yu and Jiang, 2019): A benchmark for **social media sentiment analysis**. It evaluates the alignment between textual targets and visual regions.
- **MSR-VTT** (Xu et al., 2016): A large-scale **video-to-text** benchmark with 10,000 video clips and 200,000 human-written descriptions. It evaluates video captioning through the alignment of visual dynamics and text.

Dataset	Train	Val	Test	Total
ScienceQA	12,726	4,241	4,241	21,208
MSR-VTT	6,513	497	2,990	10,000

Dataset	Positive	Neutral	Negative	Total
MVSA_Single	2,683	470	1,358	4,511
Twitter-17	2,516	728	2,728	5,972

Table 1: Statistics for multimodal datasets.

The statistics are presented in Table 1. All multimodal datasets are available at Hugging Face*.

In our experiments, there are two standardized evaluation protocols:

- **Official Protocol (ScienceQA, MSR-VTT):** To ensure fair comparisons with existing baselines, we used the official deterministic dataset splits from the original benchmarks.
- **Random Protocol (MVSA_Single, Twitter-17):** For datasets without official splits, we randomly divided the data into training, validation, and testing subsets in an 8:1:1 ratio.

Baseline Approaches. To show the efficiency and integration capability of msLoRA, we compare it against three representative PEFT baselines:

- **LoRA** (Hu et al., 2021): A unified low-rank adaptation baseline where multimodal features are naively concatenated and integrated into the LLM’s input embedding space.
- **MultiLoRA** (Wang et al., 2023): A multi-task framework utilizing parallel adapters to decouple updates; however, it lacks specific mechanisms to resolve gradient conflicts between distinct modalities.
- **AsymLoRA** (Wei et al., 2025): A multimodal variant that employs asymmetric projections to preserve distinct adaptation pathways while consolidating shared cross-modal features.

Metrics. For classification and reasoning tasks, **Accuracy** is used as the primary metric; For video understanding and captioning tasks, **CIDEr** is used to evaluate the semantic fidelity and alignment.

*https://huggingface.co/datasets/Xiuzer/multimodal_datasets/

Backbone	Method	Twitter-17						MVSA_Single					
		$r_t = 4$	$r_t = 8$	$r_t = 16$	$r_t = 32$	$r_t = 64$	$r_t = 128$	$r_t = 4$	$r_t = 8$	$r_t = 16$	$r_t = 32$	$r_t = 64$	$r_t = 128$
Llama2-7B	LoRA	0.707	0.722	0.729	0.748	0.722	0.728	0.639	0.659	0.690	0.702	0.696	0.684
	MultiLoRA	0.698	0.641	0.719	0.705	0.690	0.697	0.591	0.622	0.653	0.713	0.700	0.695
	AsymLoRA	0.676	0.725	0.737	0.725	0.716	0.720	0.573	0.641	0.649	0.692	0.690	0.679
	msLoRA	0.723	0.732	0.744	0.737	0.731	0.733	0.688	0.676	0.704	0.699	0.700	0.696
Llama2-13B	LoRA	0.519	0.720	0.733	0.729	0.731	0.707	0.687	0.680	0.694	0.684	0.702	0.695
	MultiLoRA	0.485	0.727	0.702	0.737	0.725	0.725	0.649	0.659	0.694	0.696	0.717	0.697
	AsymLoRA	0.442	0.518	0.637	0.700	0.690	0.646	0.515	0.651	0.647	0.663	0.690	0.679
	msLoRA	0.731	0.743	0.742	0.740	0.740	0.743	0.700	0.690	0.695	0.717	0.700	0.708
Qwen2.5-3B	LoRA	0.605	0.701	0.711	0.724	0.713	0.711	0.649	0.653	0.670	0.674	0.684	0.663
	MultiLoRA	0.541	0.572	0.564	0.599	0.590	0.469	0.560	0.637	0.640	0.638	0.649	0.520
	AsymLoRA	0.464	0.506	0.505	0.597	0.582	0.560	0.518	0.620	0.637	0.667	0.651	0.518
	msLoRA	0.708	0.713	0.727	0.727	0.721	0.715	0.667	0.663	0.672	0.682	0.674	0.672
Qwen2.5-7B	LoRA	0.713	0.723	0.736	0.733	0.730	0.726	0.653	0.647	0.651	0.657	0.676	0.646
	MultiLoRA	0.685	0.710	0.700	0.726	0.699	0.722	0.655	0.657	0.657	0.678	0.669	0.663
	AsymLoRA	0.545	0.652	0.705	0.703	0.705	0.689	0.518	0.618	0.632	0.672	0.658	0.669
	msLoRA	0.723	0.734	0.732	0.724	0.726	0.729	0.678	0.683	0.676	0.698	0.686	0.688
Qwen2.5-14B	LoRA	0.712	0.731	0.740	0.746	0.737	0.729	0.680	0.678	0.688	0.682	0.674	0.655
	MultiLoRA	0.562	0.652	0.714	0.725	0.737	0.724	0.620	0.694	0.663	0.678	0.690	0.684
	AsymLoRA	0.463	0.607	0.662	0.736	0.703	0.710	0.645	0.668	0.684	0.678	0.672	0.672
	msLoRA	0.727	0.741	0.747	0.737	0.737	0.726	0.692	0.700	0.692	0.677	0.663	0.684
Mistral-7B	LoRA	0.492	0.494	0.504	0.502	0.497	0.496	0.518	0.598	0.632	0.610	0.628	0.625
	MultiLoRA	0.405	0.472	0.457	0.511	0.494	0.508	0.545	0.618	0.635	0.659	0.647	0.655
	AsymLoRA	0.464	0.479	0.482	0.506	0.491	0.480	0.518	0.563	0.601	0.646	0.623	0.643
	msLoRA	0.487	0.509	0.504	0.505	0.501	0.507	0.618	0.657	0.643	0.626	0.632	0.642
		ScienceQA						MSR-VTT					
		$r_t = 4$	$r_t = 8$	$r_t = 16$	$r_t = 32$	$r_t = 64$	$r_t = 128$	$r_t = 4$	$r_t = 8$	$r_t = 16$	$r_t = 32$	$r_t = 64$	$r_t = 128$
Llama2-7B	LoRA	0.720	0.738	0.728	0.753	0.722	0.734	0.299	0.292	0.324	0.318	0.343	0.335
	MultiLoRA	0.344	0.608	0.695	0.698	0.658	0.661	0.261	0.245	0.278	0.271	0.293	0.258
	AsymLoRA	0.240	0.444	0.568	0.677	0.707	0.685	0.261	0.263	0.256	0.266	0.301	0.296
	msLoRA	0.722	0.726	0.749	0.727	0.731	0.705	0.311	0.308	0.334	0.322	0.317	0.326
Llama2-13B	LoRA	0.690	0.693	0.703	0.714	0.693	0.696	0.277	0.261	0.278	0.305	0.301	0.337
	MultiLoRA	0.198	0.351	0.431	0.656	0.672	0.570	0.183	0.243	0.272	0.294	0.271	0.279
	AsymLoRA	0.228	0.323	0.405	0.628	0.701	0.659	0.277	0.208	0.237	0.305	0.294	0.288
	msLoRA	0.724	0.667	0.685	0.698	0.712	0.701	0.290	0.330	0.313	0.306	0.315	0.316
Qwen2.5-3B	LoRA	0.344	0.478	0.539	0.732	0.717	0.709	0.307	0.284	0.260	0.322	0.316	0.321
	MultiLoRA	0.339	0.439	0.533	0.667	0.633	0.605	0.249	0.264	0.271	0.253	0.273	0.197
	AsymLoRA	0.328	0.423	0.514	0.660	0.685	0.672	0.171	0.197	0.243	0.233	0.253	0.182
	msLoRA	0.349	0.497	0.725	0.736	0.727	0.714	0.305	0.264	0.312	0.325	0.305	0.310
Qwen2.5-7B	LoRA	0.590	0.675	0.687	0.741	0.764	0.739	0.224	0.248	0.222	0.298	0.232	0.218
	MultiLoRA	0.605	0.646	0.727	0.706	0.703	0.656	0.224	0.236	0.208	0.223	0.257	0.207
	AsymLoRA	0.587	0.547	0.648	0.650	0.711	0.707	0.233	0.212	0.239	0.261	0.241	0.256
	msLoRA	0.624	0.743	0.761	0.751	0.735	0.732	0.312	0.300	0.291	0.285	0.299	0.267
Qwen2.5-14B	LoRA	0.672	0.741	0.733	0.719	0.759	0.726	0.169	0.190	0.287	0.246	0.306	0.317
	MultiLoRA	0.545	0.684	0.712	0.707	0.683	0.670	0.137	0.184	0.240	0.261	0.293	0.259
	AsymLoRA	0.421	0.543	0.532	0.642	0.707	0.700	0.152	0.178	0.253	0.297	0.274	0.295
	msLoRA	0.658	0.752	0.743	0.736	0.728	0.732	0.293	0.309	0.299	0.304	0.271	0.287
Mistral-7B	LoRA	0.319	0.342	0.344	0.330	0.315	0.327	0.206	0.294	0.281	0.290	0.285	0.285
	MultiLoRA	0.324	0.325	0.337	0.355	0.348	0.341	0.108	0.169	0.221	0.270	0.253	0.230
	AsymLoRA	0.318	0.331	0.375	0.292	0.317	0.309	0.061	0.110	0.120	0.159	0.235	0.222
	msLoRA	0.317	0.357	0.352	0.349	0.322	0.336	0.233	0.2541	0.2455	0.2671	0.2746	0.281

Table 2: Performance comparison across four multimodal datasets (CIDEr for captioning, Accuracy for others). We evaluate textual ranks $r_t \in \{4, \dots, 128\}$ on three open-source LLMs (Llama, Qwen, and Mistral). The best results for each backbone are in **bold**.

Settings. To evaluate msLoRA, we implement it across textual ranks $r_t \in \{4, \dots, 128\}$ and a uniform rank for non-textual modalities. Training spans 10 epochs for classification and reasoning tasks and 2 epochs for video understanding and captioning tasks, using a learning rate of 2×10^{-5} . The batch size (16 or 32) is adjusted based on memory limits. Visual features are extracted via pre-trained

CLIP ViT-B/32[†]; Audio features are extracted via pre-trained Wav2Vec2[‡]. All models are initialized with pre-trained weights and fine-tuned on a single RTX PRO 6000 (96GB VRAM).

[†]<https://huggingface.co/openai/clip-vit-base-patch32>

[‡]<https://huggingface.co/facebook/wav2vec2-base-960h>

In accordance with standard LoRA conventions, we target the primary weight projections within the Transformer blocks: q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, and down_proj.

5.2 Overall Performance

To evaluate the effectiveness of msLoRA across different model scales and multimodal tasks, we conduct experiments on four datasets using Llama2 (7B and 13B), Qwen2.5 (3B, 7B, and 14B), and Mistral-7B, with textual ranks $r_t \in \{4, 8, 16, 32, 64, 128\}$. Detailed results of Accuracy and CIDEr are reported in Table 2.

As shown in Table 2, msLoRA achieves strong and competitive performance across all backbones and tasks. Compared with the best result of each method across ranks, msLoRA consistently ranks among the top-performing methods. Additionally, msLoRA usually reaches its best performance at small or moderate textual ranks. This suggests that modality-specific adaptation reduces the dependence on increasing the textual LoRA rank.

Additionally, the results demonstrate that larger ranks do not always lead to better performance. Across different methods and backbones, performance is generally non-monotonic with respect to r_t , and the largest rank $r_t = 128$ is rarely optimal. This observation is consistent with the *efficiency paradox* discussed in Section 4. By using decentralized branches for modality-specific adaptation, msLoRA can exploit heterogeneous adaptation directions without relying on an excessively large shared textual rank.

Overall, msLoRA provides a good balance between multimodal performance and rank efficiency. It maintains competitive or superior performance across diverse backbones and tasks while typically achieving its best results at relatively small r_t .

5.3 Spectral and Gradient Analysis

To examine whether the modality-specific adapters learn complementary update directions, we conduct both spectral and gradient analyses on the learned msLoRA weights. These analyses are used to empirically examine the approximate modality orthogonality assumption introduced in Section 4.

Gradient Orthogonality. We first measure the cosine similarity between gradients from textual, visual, and auditory adapters. During back-propagation, we collect the gradients of the modality-specific LoRA parameters $\{A_m, B_m\}$

	txt vs. img	txt vs. aud	img vs. aud
Cosine	-0.0174	0.0214	-0.0016

Table 3: Pairwise similarity between modality features.

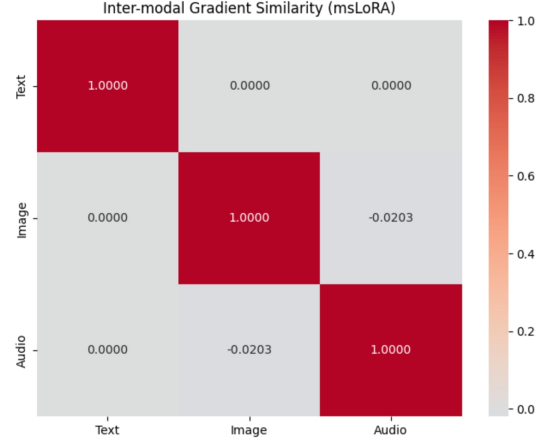


Figure 2: Heatmap of cross-modality gradient cosine similarity in msLoRA. The near-zero off-diagonal values indicate weak gradient interference.

and compute the cosine similarity between flattened gradient vectors from different modalities. As shown in Figure 2, the off-diagonal similarities are close to zero, suggesting that the modality-specific gradients are nearly orthogonal, indicating limited interference among modality updates.

Feature-Level Modality Independence. We further compute the similarity between modality features. As shown in Table 3, the pairwise cosine similarities among text, image, and audio features remain close to zero. This observation provides additional evidence that different modalities occupy weakly overlapping regions in the latent space, which supports the approximate orthogonality assumption used in Lemma 1.

Spectral Extension. We then analyze the singular value spectra of modality-specific update matrices and the fused update ΔW_{fused} . As shown in Figure 3, the text branch with $r_t = 4$ truncates after a small number of singular directions, while the image and audio branches with larger ranks preserve a wider range of spectral components. After fusion, ΔW_{fused} maintains non-negligible singular values until approximately $K = \sum_m r_m = 36$, followed by a sharp drop. This pattern suggests that modality-specific adapters contribute complementary spectral directions rather than collapsing into a single shared subspace.

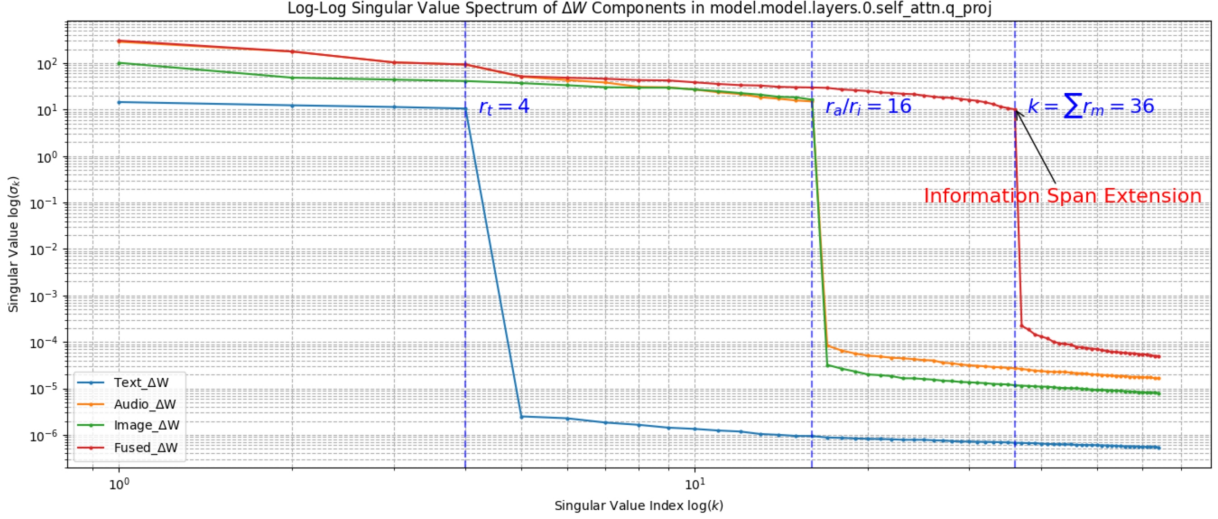


Figure 3: Log-log singular value spectrum of modality-specific ΔW components and the fused update ΔW_{fused} in the q_{proj} layer of Qwen2.5-7B on MSR-VTT dataset. The setting is $r_t = 4$, $r_i = 16$, and $r_a = 16$, leading to a cumulative rank of $K = \sum_m r_m = 36$.

	msLoRA	w/o $\mathcal{A}$	w/o $\mathcal{I}$	w/o $\mathcal{A} \& \mathcal{I}$
Llama	0.334	0.325	0.130	0.126
Qwen	0.312	0.306	0.138	0.137

Table 4: Ablation study of multimodal components on MSR-VTT. $\mathcal{A}$ and $\mathcal{I}$ denote audio and image modalities, respectively.

Overall, the gradient and spectral results provide empirical support for the **cumulative information span** analysis. The near-zero cross-modality gradient similarities indicate weak interference among modality-specific adapters, while the extended fused spectrum shows that msLoRA can preserve a broader effective update span by combining modality-specific low-rank components. These findings are consistent with the theoretical result that, under approximate modality orthogonality, the effective rank of the fused update can approach the cumulative rank $\sum_m r_m$.

5.4 Ablation Study: Effect of Multi-Modality

To quantify the contribution of each modality, we conduct an ablation study on MSR-VTT dataset using the Llama2-7B and Qwen2.5-7B backbones. We evaluated performance under three degraded configurations:

- **w/o $\mathcal{A}$:** Removing audio features to assess the impact of temporal and rhythmic cues.
- **w/o $\mathcal{I}$:** Removing visual features to evaluate the impact of spatial information.

- **w/o $\mathcal{A} \& \mathcal{I}$:** Removing non-textual modalities, reducing the system to a text-only paradigm.

As shown in Table 4, the full msLoRA model achieves the best performance on both backbones, demonstrating the benefit of integrating textual, visual, and auditory information. Removing image features leads to substantial degradation, indicating that visual cues provide the primary grounding signal for video captioning. In contrast, removing audio causes a smaller performance drop, suggesting that audio serves as a complementary modality that further enhances multimodal understanding. When both audio and image modalities are removed, the model is reduced to a text-only setting and obtains the lowest scores. These results confirm that non-textual modalities are essential for MSR-VTT and support the effectiveness of msLoRA in preserving complementary multimodal information.

5.5 Sensitivity Analysis

Effect of Asymmetric Multimodal Rank Allocation. We investigate the impact of the image rank multiplier (r_i/r_t) on CIDEr scores in the MSR-VTT dataset using Llama2-7B and Qwen2.5-7B. Here, r_i and r_t denote the image and textual ranks, respectively. As shown in Figure 4, both models achieve their best performance when the multiplier is set to 4.

The improvement from multipliers 1 to 4 suggests that non-textual modalities benefit from larger modality-specific ranks than the textual modality. This observation is consistent with our cumula-

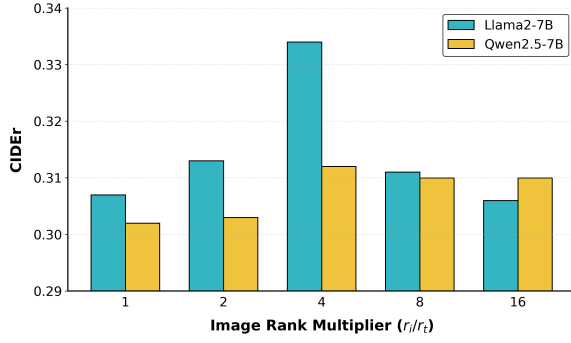


Figure 4: Effect of image rank on MSR-VTT dataset.

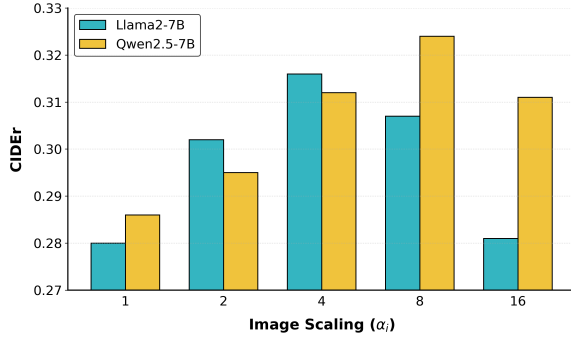


Figure 5: Impact of the image scaling (α_i) on CIDEr.

tive information span analysis, where increasing modality-specific ranks allows msLoRA to preserve more complementary spectral components. However, further increasing the multiplier to 8 or 16 leads to performance saturation or degradation, indicating that excessive ranks may introduce redundant or noisy directions. These results support the use of asymmetric rank allocation for balancing multimodal capacity and adaptation efficiency.

Effect of Image Scaling. We further investigate the sensitivity of the image scaling factor α_i on MSR-VTT dataset. As shown in Figure 5, we vary α_i over $\{1, 2, 4, 8, 16\}$ while keeping the audio scaling factor fixed.

The results show that moderate image scaling improves video captioning performance. Llama2-7B achieves its best CIDEr score when $\alpha_i = 4$, while Qwen2.5-7B reaches its peak when $\alpha_i = 8$, suggesting that stronger visual features help the model utilize spatial information better. However, further increasing α_i does not consistently improve performance and may even cause degradation, indicating that excessive visual scaling can introduce noisy or overly dominant visual signals. These results highlight the importance of balancing modality-specific contributions in msLoRA.

6 Conclusion

We propose msLoRA, a modality-specific low-rank adaptation framework for multimodal LLM fine-tuning. By assigning separate adapters to different modalities, msLoRA reduces inter-modality interference and enables heterogeneous adaptation directions to be represented through a cumulative adaptation span.

Extensive experiments across reasoning, sentiment analysis, and video captioning tasks demonstrate strong performance across multiple LLM backbones and model scales. msLoRA typically achieves its best results at small or moderate textual ranks, indicating reduced dependence on increasing the rank of a shared adapter. SVD-based spectral analysis and gradient similarity measurements further support the presence of heterogeneous modality-specific adaptation directions, consistent with our theoretical analysis. Overall, msLoRA provides an effective approach to parameter-efficient multimodal LLM adaptation.

Limitations

Our study has several limitations. First, the theoretical analysis relies on approximate orthogonality among the dominant modality-specific adaptation subspaces. When modality-specific updates are strongly aligned or highly redundant, the advantage of decentralized adapters may be reduced. Second, msLoRA introduces modality-specific hyperparameters, such as rank allocation and scaling factors, which may require additional tuning across tasks and datasets. Finally, our experiments cover four multimodal tasks and several open-source LLM backbones. Evaluating msLoRA on broader modality combinations, larger-scale instruction-tuning datasets, and additional model families remains an important direction for future work.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.62506318); Guangdong Provincial Department of Education Project (Grant No.2024KQNCX028); Scientific Research Projects for the Higher-educational Institutions (Grant No.2024312096), Education Bureau of Guangzhou Municipality; Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2025A03J3957), Education Bureau of Guangzhou Municipality.

References

- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. [A survey on evaluation of large language models](#). *ACM transactions on intelligent systems and technology*, 15(3):1–45.
- Daoquan Chen, Weicong Hong, and Xiuze Zhou. 2022. Transformer network for remaining useful life prediction of lithium-ion batteries. *Ieee Access*, 10:19621–19628.
- Kang Chen, Xiuze Zhou, Yuanguo Lin, Shibo Feng, Li Shen, and Pengcheng Wu. 2025. A survey on privacy risks and protection in large language models. *Journal of King Saud University Computer and Information Sciences*, 37(7):163.
- Yujie Ding, Shuai Jia, Tianyi Ma, Bingcheng Mao, Xiuze Zhou, Liuliu Li, and Dongming Han. 2023. Integrating stock features and global information via large language models for enhanced stock return prediction. *arXiv preprint arXiv:2310.05627*.
- Gene H Golub, Alan Hoffman, and Gilbert W Stewart. 1987. [A generalization of the eckart-young-mirsky matrix approximation theorem](#). *Linear Algebra and its applications*, 88:317–327.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. 2024. [Vtimellm: Empower llm to grasp video moments](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280.
- Yeonjoon Jung, Daehyun Ahn, Hyungjun Kim, Taesu Kim, and Eunhyeok Park. 2025. Gralora: Granular low-rank adaptation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2505.20355*.
- Ben Kabongo, Vincent Guigue, and Pirmin Lemberger. 2025. Elixir: Efficient and lightweight model for explaining recommendations. In *RecSys 2025 Workshop on Evaluating and Applying Recommender Systems with Large Language Models*. arXiv.
- Junzhuo Li, Bo Wang, Xiuze Zhou, and Xuming Hu. 2025a. Dynamic expert specialization: Towards catastrophic forgetting-free multi-domain moe adaptation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18489–18504.
- Junzhuo Li, Bo Wang, Xiuze Zhou, Peijie Jiang, Jia Liu, and Xuming Hu. 2025b. Decoding knowledge attribution in mixture-of-experts: A framework of basic-refinement collaboration and efficiency analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22431–22446.
- Yan Li, Xiangyuan Lan, Haifeng Chen, Ke Lu, and Dongmei Jiang. 2025c. [Multimodal pear chain-of-thought reasoning for multimodal sentiment analysis](#). *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(9):1–23.
- Zihao Li, Lixin Zou, Chao Ma, and Chenliang Li. 2025d. [Efficient and explainable sequential recommendation with language model](#). *Information Processing & Management*, 62(4):104122.
- Chen Lifan, Wei Ding, Jingwen Yang, Xiuze Zhou, Jingan Chen, and Fan Lin. 2026. Cogito: A cognitive agentic framework driven by dynamic graph of thoughts for financial report generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 8846–8860.
- Wuchao Liu, Wengen Li, Yu-Ping Ruan, Yulou Shu, Juntao Chen, Yina Li, Caili Yu, Yichao Zhang, Jihong Guan, and Shuigeng Zhou. 2024. [Weakly correlated multimodal sentiment analysis: New dataset and topic-oriented model](#). *IEEE Transactions on Affective Computing*, 15(4):2070–2082.
- Jie Mu, Wei Wang, Wenqi Liu, Tiantian Yan, and Guanglu Wang. 2024. [Multimodal large language model with lora fine-tuning for multimodal sentiment analysis](#). *ACM Transactions on Intelligent Systems and Technology*.
- Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. 2016. [Sentiment analysis on multi-view social data](#). In *International conference on multimedia modeling*, pages 15–27. Springer.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmlR.
- Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. [Scienceqa: A novel resource for question answering on scholarly articles](#). *International Journal on Digital Libraries*, 23(3):289–301.
- Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. 2025. [Audio-visual llm for video understanding](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4246–4255.
- Shaowen Wang, Linxi Yu, and Jian Li. 2024. [Lora-ga: Low-rank adaptation with gradient approximation](#). *Advances in Neural Information Processing Systems*, 37:54905–54931.
- Yiming Wang, Yu Lin, Xiaodong Zeng, and Guan-nan Zhang. 2023. Multilora: Democratizing lora for better multi-task learning. *arXiv preprint arXiv:2311.11501*.

- Zhaokun Wang, Jinyu Guo, Jingwen Pu, Lingfeng Chen, Hongli Pu, Jie Ou, Libo Qin, and Wenhong Tian. 2025. Noise-robustness through noise: Asymmetric lora adaption with poisoning expert. *arXiv preprint arXiv:2505.23868*.
- Xuyang Wei, Chunlin Tian, and Li Li. 2025. Asymlora: Unlocking the power of multimodal llms via asymmetric lora. In *First Workshop on Scalable Optimization for Efficient and Adaptive Foundation Models*.
- Georg Wölflein, Dyke Ferber, Daniel Truhn, Ognjen Arandjelovic, and Jakob Nikolas Kather. 2025. [Llm agents making agent tools](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26092–26130.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*.
- Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. [Msr-vtt: A large video description dataset for bridging video and language](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hiromasa Yamanishi, Ling Xiao, and Toshihiko Yamasaki. 2024. A multimodal dataset and benchmark for tourism review generation. In *The ACM RecSys Workshop on Recommenders in Tourism (RecTour 2024)*.
- Hao Yang, Yanyan Zhao, Yang Wu, Shilong Wang, Tian Zheng, Hongbo Zhang, Zongyang Ma, Wanxiang Che, Shijin Wang, Si Wei, and 1 others. 2025a. [Large language models meet text-centric multimodal sentiment analysis: A survey](#). *Science China Information Sciences*, 68(10):1–29.
- Yaming Yang, Dilxat Muhtar, Yelong Shen, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Weiwei Deng, Feng Sun, Qi Zhang, and 1 others. 2025b. [Mtl-lora: Low-rank adaptation for multi-task learning](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22010–22018.
- Jianfei Yu and Jing Jiang. 2019. [Adapting bert for target-oriented multimodal sentiment classification](#). In *Proceedings of the 28th International Joint Conference on Artificial Intelligence, IJCAI 2019*, pages 5408–5414. International Joint Conferences on Artificial Intelligence.
- Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, and 1 others. 2025. [Vidorefer suite: Advancing spatial-temporal object understanding with video llm](#). In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18970–18980.
- Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. 2023. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pages 41753–41769. PMLR.