Parameter-Efficient Adaptation for MLLMs via Implicit Modality Decomposition

Supplementary Material

A. Algorithm

A.1. Modality-Specific Decoupling Gradient

As described in the main paper, instead of aggregating the Modality-Specific Decoupling constraints $\mathcal{L}_{\text{TS}}$ and $\mathcal{L}_{\text{OS}}$ into the overall loss, we compute the parameter-wise gradients and inject them into the backward pass, as shown in Eq.(6). We now derive the closed-form expressions of the gradient term $\mathbf{G}_{\mathbf{A}}^{\text{MSD}}$.

Let $\mathbf{M}_t, \mathbf{M}_o, \mathbf{M}_s \in \{0,1\}^{r\times d}$ be the binary masks for text-specific, non-text-specific and shared parameters, the mask of parameters accessible to text is defined as $\mathbf{M} = \mathbf{M}_t + \mathbf{M}_s$. For a text-only input $\mathbf{X}_t \in \mathbb{R}^{d\times L}$, the text-decoupling loss from Eq.(5) can be written as:

$$\mathcal{L}_{\text{TS}} = ||\tilde{\mathbf{z}}_t - \mathbf{z}_t||_2^2 = ||(\mathbf{A}\odot(\mathbf{1}-\mathbf{M}))\mathbf{X}_t)||_F^2. \tag{13}$$

Since $\mathbf{M}_t + \mathbf{M}_s + \mathbf{M}_o = \mathbf{1}$, we have $\mathbf{1} - \mathbf{M} = \mathbf{M}_o$. Hence Eq. (13) simplifies to the explicit form:

$$\mathcal{L}_{\text{TS}} = ||(\mathbf{A}\odot\mathbf{M}_o)\mathbf{X}_t)||_F^2, \tag{14}$$

which penalizes the contribution of non-text-specific parameters when processing text-only inputs.

Differentiating Eq. (14) with respect to $\mathbf{A}$ yields:

$$\nabla_{\mathbf{A}}\mathcal{L}_{\text{TS}} = 2((\mathbf{A}\odot\mathbf{M}_o)\mathbf{X}_t)\mathbf{X}_t^T \odot \mathbf{M}_o. \tag{15}$$

Similarly, we derive the expression of non-text decoupling gradient:

$$\nabla_{\mathbf{A}}\mathcal{L}_{\text{OS}} = 2((\mathbf{A}\odot\mathbf{M}_t)\mathbf{X}_o)\mathbf{X}_o^T \odot \mathbf{M}_t. \tag{16}$$

Combining Eq. (15) and Eq. (16) gives the final closed form used during backward injection:

$$\begin{aligned}\mathbf{G}_{\mathbf{A}}^{\text{MSD}} = 2\Big[&\big((\mathbf{A}\odot\mathbf{M}_o)\mathbf{X}_t\big)\mathbf{X}_t^T\odot\mathbf{M}_o \\ &+\big((\mathbf{A}\odot\mathbf{M}_t)\mathbf{X}_o\big)\mathbf{X}_o^T\odot\mathbf{M}_t\Big].\end{aligned} \tag{17}$$

A.2. Modality-Agnostic Alignment Gradient

Next, we provide the detailed gradient derivation for the Modality-Agnostic Alignment Gradient $\mathbf{G}_{\mathbf{A}}^{\text{MA}}$, corresponding to Eq.(11) in the main paper. As introduced in Eq.(10), the Modality-Agnostic Alignment Constraint is formulated as a cosine-based alignment term modulated by centralization factors. The centralization factors quantify semantic concentration and serve only as weighting coefficients, not intended to directly alter the optimization direction. Therefore, when computing the gradient, we treat them as a constant scalar weight w. With this simplification, the gradient of the alignment loss is written as:

$$\nabla_{\mathbf{A}}\mathcal{L}_{\text{MA}} = -w\cdot\rho,\ \ \rho = \frac{\bar{\mathbf{s}}_t^T\bar{\mathbf{s}}_o}{||\bar{\mathbf{s}}_t||\,||\bar{\mathbf{s}}_o||}. \tag{18}$$

To compute the gradient with respect to $\mathbf{A}$, we first obtain the partial derivatives of the cosine similarity ρ with respect to the mean semantic representations of text and non-text modalities:

$$\frac{\partial\rho}{\partial\bar{\mathbf{s}}_t} = \frac{\bar{\mathbf{s}}_o}{||\bar{\mathbf{s}}_t||\,||\bar{\mathbf{s}}_o||} - \rho\frac{\bar{\mathbf{s}}_t}{||\bar{\mathbf{s}}_t||^2}, \tag{19}$$

$$\frac{\partial\rho}{\partial\bar{\mathbf{s}}_o} = \frac{\bar{\mathbf{s}}_t}{||\bar{\mathbf{s}}_t||\,||\bar{\mathbf{s}}_o||} - \rho\frac{\bar{\mathbf{s}}_o}{||\bar{\mathbf{s}}_o||^2}. \tag{20}$$

The summary representations $\bar{\mathbf{s}}_t$ and $\bar{\mathbf{s}}_o$ are obtained by applying the shared-modality mask M_s to the LoRA matrix $\mathbf{A}$, followed by a linear transformation using the mean input vectors:

$$\bar{\mathbf{s}}_t = (A\odot M_s)v_t,\ \ \bar{\mathbf{s}}_o = (A\odot M_s)v_o, \tag{21}$$

Algorithm 1: IMoD Training Algorithm

Input: Training data $(\mathbf{X}_t, \mathbf{X}_o)$; pre-trained weight $\mathbf{W}$; trainable LoRA matrices $\mathbf{A}, \mathbf{B}$

```
/* Forward pass of a single IMoD module */
Function Forward(X_t, X_o):
    // Compute output same as LoRA
    X ← [X_o; X_t];
    Ouput ← (W + BA)X;
    // Compute gradient of two constraint
    G_A^MSD ← Eq. (17);
    G_A^MA ← Eq. (22);
    return Output;
End
/* Backward pass of a single IMoD module */
Function Backward(G_output):
    // Compute gradient same as LoRA
    G_B ← G_output X^T A^T;
    G_A ← (B^T G_output) X^T;
    // Add gradient of two constraints
    G_A ← G_A + λ1 G_A^MSD + λ2 G_A^MA
End
```

Methods	Audio-Visual-Text		Visual-Text				Audio-Text	
	MUSIC-AVQA	AVE	MME_{percep}	MMBench	POPE	SEED-Bench	MMAU	Air-Bench
LoRA	72.83	72.13	1062.34	57.89	81.17	55.25	50.15	44.55
Multiple LoRA	72.71	72.11	1103.28	57.01	80.96	55.13	50.45	41.13
LoRAMoE	73.48	72.83	1157.39	57.29	81.29	56.39	50.75	42.99
DoRA	73.29	72.91	1024.42	56.19	80.75	55.03	52.55	44.11
HydraLoRA	73.14	72.59	1098.25	56.42	81.34	54.67	53.45	43.94
Uni-modal LoRA	73.62	73.14	1189.47	57.39	81.12	56.21	54.05	47.01
MokA	75.26	74.48	1292.37	59.06	82.33	58.10	55.26	49.17
IMoD (Ours)	**77.34**	**75.67**	**1323.39**	**60.18**	**82.87**	**59.77**	**56.98**	**50.43**

Table A. **Comparison with State-of-the-Art Methods on Qwen2**. We consistently outperform prior approaches on all multimodal scenarios and benchmarks.

Here, $v_t = \frac{1}{L}\sum_i x_{t,i}$ and $v_o = \frac{1}{L}\sum_i x_{o,i}$ denote the modality-wise mean input features for text and non-text tokens, respectively.

By combining the above terms using the chain rule, we obtain the final gradient of the Modality-Agnostic Alignment Constraint with respect to $\mathbf{A}$:

$$\mathbf{G}_{\mathbf{A}}^{\text{MA}} = -w\left[\left(\frac{\bar{\mathbf{s}}_o}{\|\bar{\mathbf{s}}_t\|\|\bar{\mathbf{s}}_o\|} - \rho\frac{\bar{\mathbf{s}}_t}{\|\bar{\mathbf{s}}_t\|^2}\right)v_t^T + \left(\frac{\bar{\mathbf{s}}_t}{\|\bar{\mathbf{s}}_t\|\|\bar{\mathbf{s}}_o\|} - \rho\frac{\bar{\mathbf{s}}_o}{\|\bar{\mathbf{s}}_o\|^2}\right)v_o^T\right] \odot M_s \quad (22)$$

A.3. Overall Training Algorithm

This section presents the overall training pipeline of IMoD, integrating the two proposed parameter-level constraints into standard LoRA fine-tuning. As shown in Algorithm 1, the forward pass performs the same computation as LoRA, using the combined update matrix to obtain the module output. During this forward pass, we additionally compute the gradients associated with the Modality-Specific Decoupling (MSD) and Modality-Agnostic Alignment (MA) constraints using their closed-form expressions, without injecting them into the forward outputs.

In the backward stage, IMoD preserves the original LoRA gradient computation for matrix $\mathbf{A}$ and $\mathbf{B}$. After the standard LoRA gradients are computed, we add the constraint-induced gradients to the update of $\mathbf{A}$, scaled by λ_1 and λ_2. This procedure ensures that the task loss continues to drive the main optimization while the constraints provide fine-grained parameter-level guidance toward modality specialization and modality invariance, respectively. Importantly, this design maintains full compatibility with inference-time LoRA weight merging and introduces no additional trainable parameters.

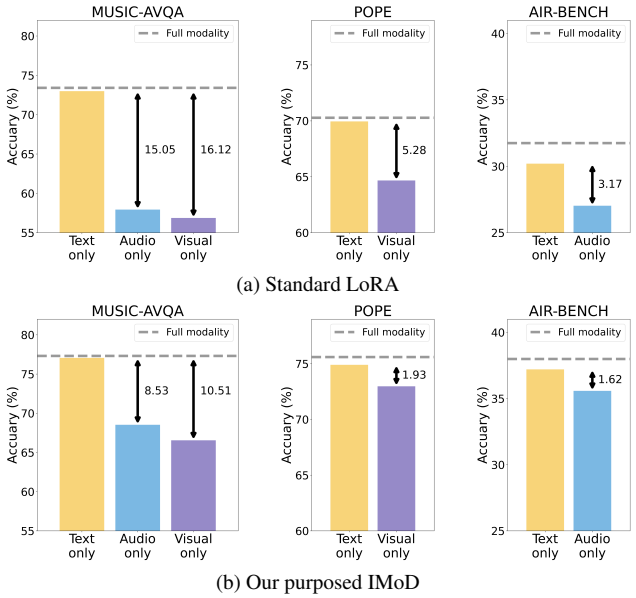


Figure A. **Verification of Balanced Modality.** Full modality is the regular case where all multimodal tokens are processed by the fine-tuned module, whereas text/visual/audio-only represents only the single selected modality is passed through the fine-tuned module. The black arrows indicate the performance gap between the text-only and visual/audio-only settings, a larger gap reflects a more severe modality imbalance. The results show our method significantly alleviate modality imbalance.

B. Experiments

B.1. Additional Experiments on Qwen2

To further validate the generality and effectiveness of IMoD, we conducted supplementary experiments using the Qwen2-7B-Instruct backbone across multiple multimodal scenarios, including audio-visual-text, visual-text, and speech-text tasks. As summarized in Tab. A, IMoD consistently achieves state-of-the-art across all evaluated benchmarks without increasing the number of trainable pa-

Text/Non-text	Exist	Count	Location	Compare	Temporal
0.5	81.56	**81.28**	69.47	59.73	70.15
1.0	81.96	80.58	**70.44**	**60.18**	**70.51**
2.0	**82.06**	80.11	68.28	59.82	**70.51**

Table B. Impact of mask ratios on different tasks.

rameters or incurring additional inference cost. This result further highlights the broad applicability of our approach to multiple LLM backbones.

B.2. Ablation of Mask Ratios Across Tasks

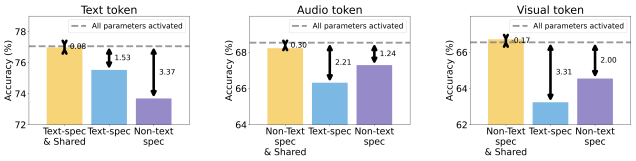
In this subsection, we investigate how mask ratios affect model performance. As shown in Tab. B, the mask ratios define the pre-allocated proportions of text-specific and non-text-specific parameters, reflecting the model's capacity to learn textual versus non-textual information. Since different tasks rely on modalities to varying extents, the optimal mask ratio is task-dependent. We therefore conduct experiments across multiple tasks with different mask ratios. The results demonstrate that the optimal ratios indeed vary across tasks, while the overall performance remains stable, indicating that IMoD can flexibly adapt to different modality requirements.

B.3. Effect of Balancing Modality

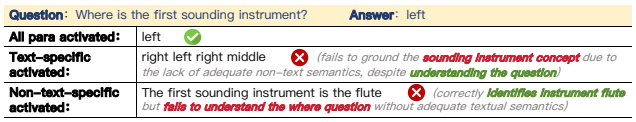
To validate the effectiveness of our method in balancing the contribution of different modalities, we conduct partial-modality inference experiments following the protocol established in prior works. For all experiments, the model is trained with full multimodal tokens processed through the LoRA module under the standard training pipeline. At inference, we selectively feed tokens from a single modality into the LoRA module at the first generation step and evaluate the performance. As shown in Fig. Aa, models fine-tuned with standard LoRA achieve nearly identical performance when using only text tokens compared to all modalities across audio-visual-text, visual-text, and speech-text scenarios. However, inference with only non-text tokens leads to a severe performance degradation, indicating that shared LoRA parameters are overly dominated by text tokens, which reduces the effective utilization of non-text modalities. In contrast, Fig. Ab demonstrates that our proposed IMoD considerably narrows the performance gap between text-only and non-text-only inference. This result confirms that IMoD effectively facilitates more balanced and sufficient utilization of different modalities.

B.4. Verification of Semantic Specialization

We evaluate semantic specialization with partial-activation inference as shown in Fig. Ba. Only tokens from one target modality are forwarded through the trainable IMoD parameters under partial activation, while tokens from all other



(a) Quantitative results of selective IMoD parameter activation.



(b) Qualitative example of selective IMoD parameter activation.

Figure B. Semantic specialization of IMoD parameter groups.

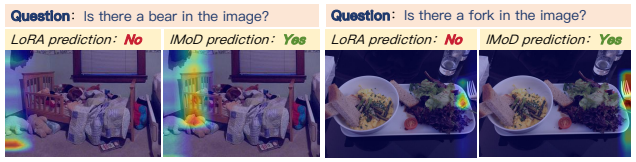


Figure C. Qualitative comparison between LoRA and IMoD.

modalities are processed exclusively with the frozen parameters. The quantitative results indicate that text-specific parameters mainly encode textual semantics, while non-text-specific parameters capture visual and audio semantics. Qualitative examples, where all tokens pass through trainable parameters under partial activation, are presented in Fig. Bb to show semantic differences.

B.5. Visualization of Attentions

We visualize the attention from the target word to visual tokens, overlaid on the image. As shown in Fig. C, LoRA exhibits misaligned attention, while IMoD focuses on relevant regions and yields correct prediction.