LoRASculpt: Sculpting LoRA for Harmonizing General and Specialized Knowledge in Multimodal Large Language Models

Supplementary Material

A. Proof of Theorem 3.1

Theorem 3.1. *Let $B \in \mathbb{R}^{p\times r}$ and $A \in \mathbb{R}^{r\times q}$ be two low rank matrices in LoRA, then the expected sparsity of the product matrix $BA \in \mathbb{R}^{p\times q}$ is given by:*

$$\mathbb{E}\left[s_{BA}\right] = 1 - (1 - s_B s_A)^r. \tag{12}$$

Proof. We aim to determine the expected proportion of non-zero elements in the product matrix $BA \in \mathbb{R}^{p\times q}$. The element in the i-th row and j-th column of BA is given by

$$(BA)_{ij} = \sum_{k=1}^{r} B_{ik} A_{kj}. \tag{13}$$

We will prove a stronger conclusion: we assume that all elements in A and B are nonnegative. This assumption increases the number of nonzero elements in BA, making it more challenging to ensure the sparsity of BA.

Then an element $(BA)_{ij}$ is non-zero if and only if there exists at least one $k \in \{1, 2, \ldots, r\}$ such that both B_{ik} and A_{kj} are non-zero. For each k, the probability that B_{ik} is non-zero is s_B, and the probability that A_{kj} is non-zero is s_A. Since the positions of non-zero elements in B and A are independently and randomly distributed, the probability that both B_{ik} and A_{kj} are non-zero is

$$\mathbb{P}\left(B_{ik} \neq 0 \text{ and } A_{kj} \neq 0\right) = s_B s_A. \tag{14}$$

Therefore, the probability that $B_{ik}A_{kj} = 0$ is

$$\mathbb{P}\left(B_{ik}A_{kj} = 0\right) = 1 - s_B s_A. \tag{15}$$

Assuming independence across different k, the probability that all terms $B_{ik}A_{kj}$ are zero is

$$\begin{aligned}\mathbb{P}\left(\bigcap_{k=1}^{r}\{B_{ik}A_{kj} = 0\}\right) &= \prod_{k=1}^{r}\mathbb{P}\left(B_{ik}A_{kj} = 0\right) \\ &= \left(1 - s_B s_A\right)^r.\end{aligned} \tag{16}$$

Thus, the probability that $(BA)_{ij}$ is non-zero is

$$\begin{aligned}\mathbb{P}\left((BA)_{ij} \neq 0\right) &= 1 - \mathbb{P}\left((BA)_{ij} = 0\right) \\ &= 1 - \left(1 - s_B s_A\right)^r.\end{aligned} \tag{17}$$

Since there are $p \times q$ elements in BA, the expected number of non-zero elements is

$$\mathbb{E}\left[N_{BA}\right] = pq\left[1 - \left(1 - s_B s_A\right)^r\right], \tag{18}$$

where N_{BA} denotes the number of non-zero elements in BA.

The expected sparsity of BA is then

$$\mathbb{E}\left[s_{BA}\right] = \frac{\mathbb{E}\left[N_{BA}\right]}{pq} = 1 - \left(1 - s_B s_A\right)^r. \tag{19}$$

The proof of Theorem 3.1 is finished. □

B. Proof of Theorem 3.2

Theorem 3.2. *Let $B \in \mathbb{R}^{p\times r}$ and $A \in \mathbb{R}^{r\times q}$ be two low rank matrices in LoRA, where the sparsity of B is s_B and the sparsity of A is s_A. Define $C = BA$, with sparsity s_C. Then, for any $\delta > 0$:*

$$\mathbb{P}\left(|s_C - \mathbb{E}[s_C]| \geq \delta\right) \leq 2\exp\left(-\frac{2\delta^2 pq}{r(p+q)}\right), \tag{20}$$

where the expected sparsity $\mathbb{E}[s_C]$ is given by Theorem 3.1

Proof. We aim to apply McDiarmid's inequality to the total number of nonzero entries N in C.

McDiarmid's Inequality states that if $X_1, X_2, \ldots, X_n$ are independent random variables taking values in a set $\mathcal{X}$, and $f : \mathcal{X}^n \to \mathbb{R}$ satisfies the bounded differences condition: for all i and all $x_1, \ldots, x_n, x_i' \in \mathcal{X}$,

$$|f(x_1, \ldots, x_i, \ldots, x_n) - f(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i,$$

then for all $\epsilon > 0$,

$$\mathbb{P}\left(f(X_1, \ldots, X_n) - \mathbb{E}[f] \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n} c_i^2}\right),$$

and similarly for $\mathbb{P}\left(\mathbb{E}[f] - f(X_1, \ldots, X_n) \geq \epsilon\right)$.

In our context, consider the function f representing the total number of nonzero entries in C:

$$N = \sum_{i=1}^{p}\sum_{j=1}^{q} X_{ij}, \tag{21}$$

where X_{ij} is the indicator variable:

$$X_{ij} = \begin{cases} 1, & \text{if } C_{ij} \neq 0, \\ 0, & \text{if } C_{ij} = 0. \end{cases} \tag{22}$$

Each C_{ij} depends on the random variables $\{B_{ik}, A_{kj}\}_{k=1}^{r}$. The variables B_{ik} and A_{kj} are independent and affect N through C_{ij}.

We have the bounded differences:

Effect of changing B_{ik}: Changing B_{ik} can affect all C_{ij} where $j = 1, \ldots, q$. The maximum change in N due to changing B_{ik} is $c_{B_{ik}} = q$.

Effect of changing A_{kj}: Changing A_{kj} can affect all C_{ij} where $i = 1, \ldots, p$. The maximum change in N due to changing A_{kj} is $c_{A_{kj}} = p$.

Therefore, the sum of the squares of the bounded differences is:

$$\sum_{i,k} c_{B_{ik}}^2 + \sum_{k,j} c_{A_{kj}}^2 = pr \cdot q^2 + rq \cdot p^2 = rpq(p+q). \tag{23}$$

Applying McDiarmid's inequality, for any $\epsilon > 0$:

$$\mathbb{P}\left(N - \mathbb{E}[N] \geq \epsilon\right) \leq \exp\left(-\frac{2\epsilon^2}{rpq(p+q)}\right), \tag{24}$$

and similarly for $\mathbb{P}\left(\mathbb{E}[N] - N \geq \epsilon\right)$. Therefore,

$$\mathbb{P}\left(|N - \mathbb{E}[N]| \geq \epsilon\right) \leq 2\exp\left(-\frac{2\epsilon^2}{rpq(p+q)}\right). \tag{25}$$

Since $s_C = \dfrac{N}{pq}$, we have:

$$|s_C - \mathbb{E}[s_C]| = \frac{|N - \mathbb{E}[N]|}{pq}. \tag{26}$$

Let $\delta = \dfrac{\epsilon}{pq}$, so $\epsilon = \delta pq$. Substituting back into the inequality:

$$\begin{aligned}\mathbb{P}\left(|s_C - \mathbb{E}[s_C]| \geq \delta\right) &\leq 2\exp\left(-\frac{2(\delta pq)^2}{rpq(p+q)}\right) \\ &= 2\exp\left(-\frac{2\delta^2 pq}{r(p+q)}\right).\end{aligned} \tag{27}$$

The proof of Theorem 3.2 is finished.

□

C. Proof of Theorem D.1

Theorem D.1. *Consider matrices $A \in \mathbb{R}^{r \times q}$ and $B \in \mathbb{R}^{p \times r}$, where each row of B and each column of A exhibit uniform sparsity internally but vary across rows and columns, respectively, with average sparsities s_A and s_B. Then, the expected proportion $\mathbb{E}[s_C]$ of nonzero entries in the product matrix $C = BA$ satisfies:*

$$\mathbb{E}[s_C] \leq 1 - (1 - s_A s_B)^r. \tag{28}$$

Proof. Consider any entry c_{ij} of the matrix $C = BA$, which is computed as:

$$c_{ij} = \sum_{k=1}^{r} b_{ik} a_{kj}. \tag{29}$$

To determine the probability that c_{ij} is nonzero, we analyze the sparsity of b_{ik} and a_{kj}.

For fixed i and j, we define: s_{B_i} is the sparsity of the i-th row of B; s_{A_j} is the sparsity of the j-th column of A,.

For b_{ik} and a_{kj}, we have:

$$\mathbb{P}(b_{ik} \neq 0) = s_{B_i}, \quad \mathbb{P}(a_{kj} \neq 0) = s_{A_j}. \tag{30}$$

Since the positions of nonzero elements within the i-th row of B and the j-th column of A are independently and uniformly distributed, the events that b_{ik} and a_{kj} are nonzero are independent for each k. Therefore, the probability that both b_{ik} and a_{kj} are nonzero is:

$$\mathbb{P}(b_{ik} \neq 0 \text{ and } a_{kj} \neq 0) = s_{B_i} s_{A_j}. \tag{31}$$

Same as the proof for Theorem 3.1, we prove a stronger conclusion by assuming that all elements in A and B are nonnegative. Thus, for $c_{ij} = 0$, it must hold that for all $k = 1, 2, \ldots, r$, either $b_{ik} = 0$ or $a_{kj} = 0$. Consequently, the probability that $c_{ij} = 0$ is:

$$\begin{aligned}\mathbb{P}(c_{ij} = 0) &= \prod_{k=1}^{r} \left[1 - \mathbb{P}(b_{ik} \neq 0 \text{ and } a_{kj} \neq 0)\right] \\ &= \left(1 - s_{B_i} s_{A_j}\right)^r.\end{aligned} \tag{32}$$

Thus, the probability that c_{ij} is nonzero is:

$$\begin{aligned}\mathbb{P}c_{ij} \neq 0) &= 1 - \mathbb{P}(c_{ij} = 0) \\ &= 1 - \left(1 - s_{B_i} s_{A_j}\right)^r.\end{aligned} \tag{33}$$

Therefore, the expected proportion of nonzero entries in C is:

$$\mathbb{E}[s_C] = \frac{1}{pq} \sum_{i=1}^{p} \sum_{j=1}^{q} \left[1 - \left(1 - s_{B_i} s_{A_j}\right)^r\right]. \tag{34}$$

Note that for $x \in [0, 1]$ and $r \geq 1$, the function $f(x) = (1 - x)^r$ is convex. According to Jensen's Inequality, for a convex function f and a random variable X, we have:

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]). \tag{35}$$

In our case, let the random variables be $X_{ij} = s_{B_i} s_{A_j}$, then:

$$\begin{aligned}\mathbb{E}[X] &= \frac{1}{pq} \sum_{i=1}^{p} \sum_{j=1}^{q} X_{ij} \\ &= \left(\frac{1}{p} \sum_{i=1}^{p} s_{B_i}\right) \left(\frac{1}{q} \sum_{j=1}^{q} s_{A_j}\right) \\ &= s_B s_A.\end{aligned} \tag{36}$$

Applying Jensen's Inequality, we obtain:

$$\frac{1}{pq} \sum_{i=1}^{p} \sum_{j=1}^{q} \left(1 - s_{B_i} s_{A_j}\right)^r \geq \left(1 - s_B s_A\right)^r. \tag{37}$$

That is:

$$\begin{aligned}1 - \mathbb{E}[s_C] &= \frac{1}{pq} \sum_{i=1}^{p} \sum_{j=1}^{q} \left(1 - s_{B_i} s_{A_j}\right)^r \\ &\geq \left(1 - s_B s_A\right)^r.\end{aligned} \tag{38}$$

From the inequality above, we have:

$$\mathbb{E}[s_C] \leq 1 - \left(1 - s_B s_A\right)^r. \tag{39}$$

The proof of Theorem 3.3 is finished.

□

D. Proof of LoRASculpt Sparsity Guarantee

We first demonstrate that incorporating Knowledge-Guided Regularization impacts the sparsity structure of the low-rank LoRA matrices. Specifically, each row of B maintains uniform sparsity, though different rows have varied sparsity levels; similarly, each column of A has consistent sparsity within itself, while sparsity varies across columns. The overall sparsity of the two low-rank matrices remains at s_B and s_A following one-shot pruning. For the partial derivative of the (i, j)-th element of the delta weight BA, we have the following expression:

$$\frac{\partial \mathcal{L}_{CMR}^2}{\partial (BA)_{ij}} = 2 \cdot M_{ij}^2 \cdot \sum_k B_{ik} A_{kj}, \tag{40}$$

This indicates that the penalty on the (i, j)-th position in the delta weight BA affects the i-th row of B and the j-th column of A . Consequently, B is constrained by rows, and A by columns, resulting in varying sparsity across the rows of B and the columns of A. Under this condition, the following theorem holds:

Theorem D.1. *Consider matrices $A \in \mathbb{R}^{r \times q}$ and $B \in \mathbb{R}^{p \times r}$, where each row of B and each column of A exhibit uniform sparsity internally but vary across rows and columns, respectively, with average sparsities s_A and s_B. Then, the expected proportion $\mathbb{E}[s_C]$ of nonzero entries in the product matrix $C = BA$ satisfies:*

$$\mathbb{E}[s_C] \leq 1 - (1 - s_A s_B)^r. \tag{41}$$

Proof. See Appendix C. □

Despite the non-uniform sparsity of matrices B and A across rows and columns, where different rows of B and different columns of A exhibit varied distributions, we can still assume the independence of updates across all elements. This does not hinder the application of McDiarmid's inequality, thereby allowing us to obtain the previously established error bounds in Theorem 3.2. Thus, we have established the sparsity guarantees of LoRASculpt.

E. Algorithm of LoRASculpt

The algorithm is outlined in Algorithm 1. Please refer to Sec. 3.2 for more details.

```
Algorithm 1: LoRASculpt
Input: Training Steps $T$, Warmup Steps $T_{\text{warmup}}$, Training
       data $\mathcal{D}_{\text{tr}}$, Sparsity Ratio $s_A$, $s_B$ , Number of Layer
       in LLM and Connector $L_{LLM}$, $L_{Con}$, Option of
       whether training Connector with LoRA Flag$_{Con}$.
Output: Final LoRA weights.
$\mathcal{L}_{CMR}^{\text{LLM}} \leftarrow 0, \quad \mathcal{L}_{CMR}^{\text{Con}} \leftarrow 0$ ;
$S \leftarrow \psi(W) = \left| 1/\log\left(\frac{|W|}{\|W\|_2} + \epsilon\right) \right|$ ;   ▷ Eq. (6)
$M \leftarrow \tanh(\omega \odot S)$ ;   ▷ Eq. (7)
for $t = 1, 2, \ldots, T$ do
    Sample a batch $(x^{\text{vision}}, x^{\text{text}}, y)$ in $\mathcal{D}_{\text{tr}}$ ;
    if $t \geq T_{\text{warmup}}$ then
        if $t = T_{\text{warmup}}$ then
            $M_A \leftarrow \text{Mask}(A, s_A)$ ;
            $M_B \leftarrow \text{Mask}(B, s_B)$ ;   ▷ Eq. (3)
        end
        $A \leftarrow M_A \odot A$ ;
        $B \leftarrow M_B \odot B$ ;   ▷ Eq. (2)
    end
    $h^{\text{vision}} = \varphi_{Con} \circ \varphi_{Vis}(x^{\text{vision}}), h^{\text{text}} = \text{Tokenize}(x^{\text{text}})$ ;
    $\mathcal{L}_{Task} \leftarrow \mathcal{L}_{CE}\left(\Phi\left[h^{\text{vision}}, h^{\text{text}}\right], y\right)$ ;
    for $l = 1, 2, \ldots, L_{LLM}$ do
        $\mathcal{L}_{CMR}^{\text{LLM}} \leftarrow \mathcal{L}_{CMR}^{\text{LLM}} + \|M_l \odot (B_l A_l)\|_F$ ;   ▷ Eq. (8)
    end
    if Flag$_{Con}$ $=$ True then
        for $\tilde{l} = 1, 2, \ldots, L_{Con}$ do
            $\mathcal{L}_{CMR}^{\text{Con}} \leftarrow \mathcal{L}_{CMR}^{\text{Con}} + \|M_{\tilde{l}} \odot (B_{\tilde{l}} A_{\tilde{l}})\|_1$ ;
                ▷ Eq. (10)
        end
    end
    $\mathcal{L} = \mathcal{L}_{Task} + \alpha \cdot \mathcal{L}_{CMR}^{\text{LLM}} + \beta \cdot \mathcal{L}_{CMR}^{\text{Con}}$ ;   ▷ Eq. (11)
    Update low-rank adapters to minimize $\mathcal{L}$ ;
end
return Fine-tuned LoRA in $\Phi$ (and $\varphi_{Con}$)
```

F. Addition Evaluation Details

Details of Compared Baselines

(a) LoRA [ICLR'22] [17]: Introduces low-rank adapters to efficiently fine-tune large models.
(b) DoRA [ICML'24] [39]: Enhances the learning capacity and training stability of LoRA by decomposing weights into magnitude and direction.
(c) Orth-Reg [ECCV'24] [18]: Adds an orthogonal regularization with a hyperparameter (*i.e.*, 1e-3) to LoRA weights, encouraging fine-tuned features to be orthogonal to pretrained features to preserve model generalization. For fair comparison and due to resource constraints, the component that involves multiple LoRA modules is excluded.
(d) L2-Regularization [PNAS'17] [28]: Apply L_2 regularization with a hyperparameter (*i.e.*, 1e-3) to the LoRA weights, guiding the fine-tuned model closer to the pretrained model thus reducing forgetting.
(e) DARE [ICML'24] [73]: Parameters from the fine-tuned LoRA weights are randomly selected and re-scaled to mitigate knowledge conflict of the target task and other tasks.
(f) Model Tailor [ICML'24] [84]: Retains pretrained parameters while selectively replacing a small portion (*i.e.*, 10%) of fine-tuned parameters, guided by salience and sensitivity analysis.

Evaluation Metric.

To evaluate the performance of MLLMs in general and specialized knowledge, we compute the source performance (denotes by *Source*) and target performance (denotes by *Target*):

$$\textit{Source} = \frac{1}{|\mathcal{D}|} \sum_{i}^{|\mathcal{D}|} \text{Score}(\mathcal{D}_i), \quad \textit{Target} = \text{Score}(\mathcal{T}). \tag{42}$$

where $\text{Score}(\cdot)$ denotes the evaluation metric for different datasets, which is set to Accuracy and CIDEr for VQA and Captioning tasks, respectively. Here, $\mathcal{D} = \{\mathcal{D}_i\}_{i=1}^{|\mathcal{D}|}$ represents the datasets used to evaluate general knowledge, and $\mathcal{T}$ denotes the downstream task dataset. We use the average score of *Source* and *Target*, denoted as *Avg* to measure the overall capability of the MLLM.

G. Ablation Study of β

β controls the sparsity strength for MLLM connector in Eq. (11). Since the connector plays a crucial role in modality alignment, adopting a high sparsity level could lead to performance degradation on downstream tasks (denoted by *Target* in Tab. I). Selecting an appropriate β to sparsify the connector can achieve a balance between *Source* and *Target*.

β	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
Source	60.19	58.58	59.73	59.55	59.13
Target	80.10	79.99	84.02	85.34	85.01
Avg	70.15	69.29	71.87	**72.45**	72.07

Table I. **Ablation Study of** β, which represents the intensity of sparsity applied to the MLLM connector. When set to 10^{-5}, the optimal *Avg* is achieved.