O-LoRA

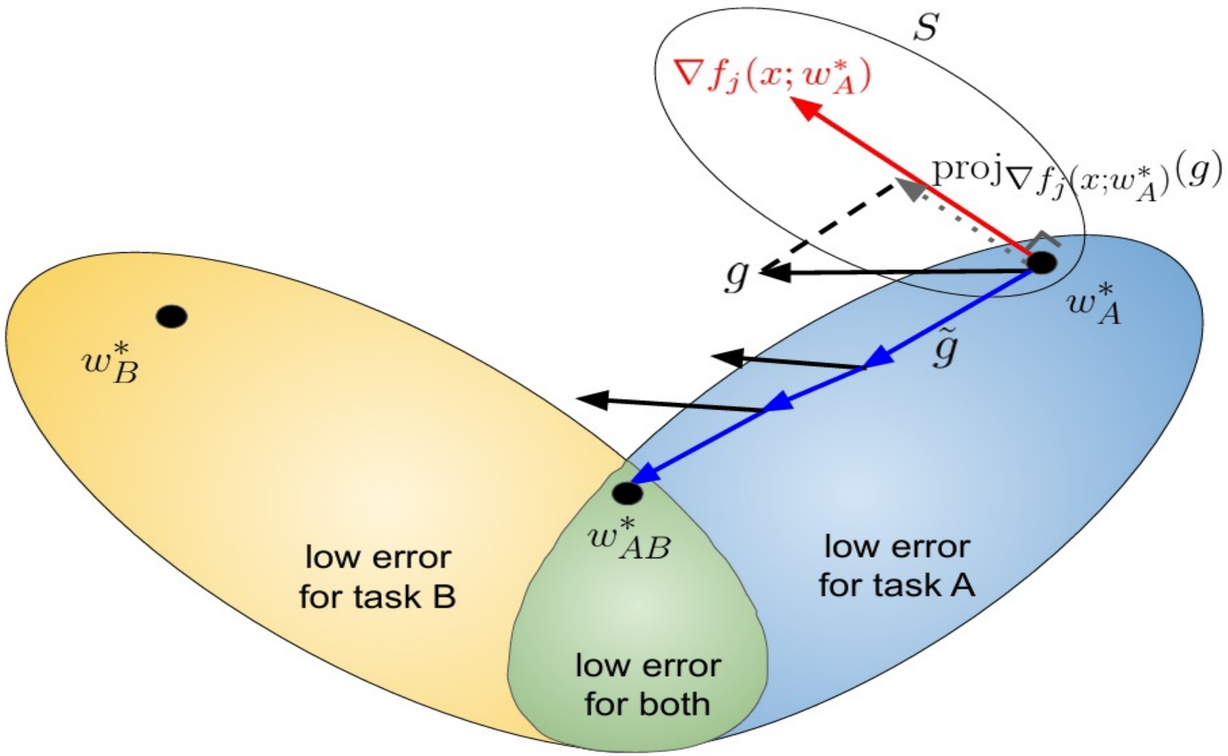
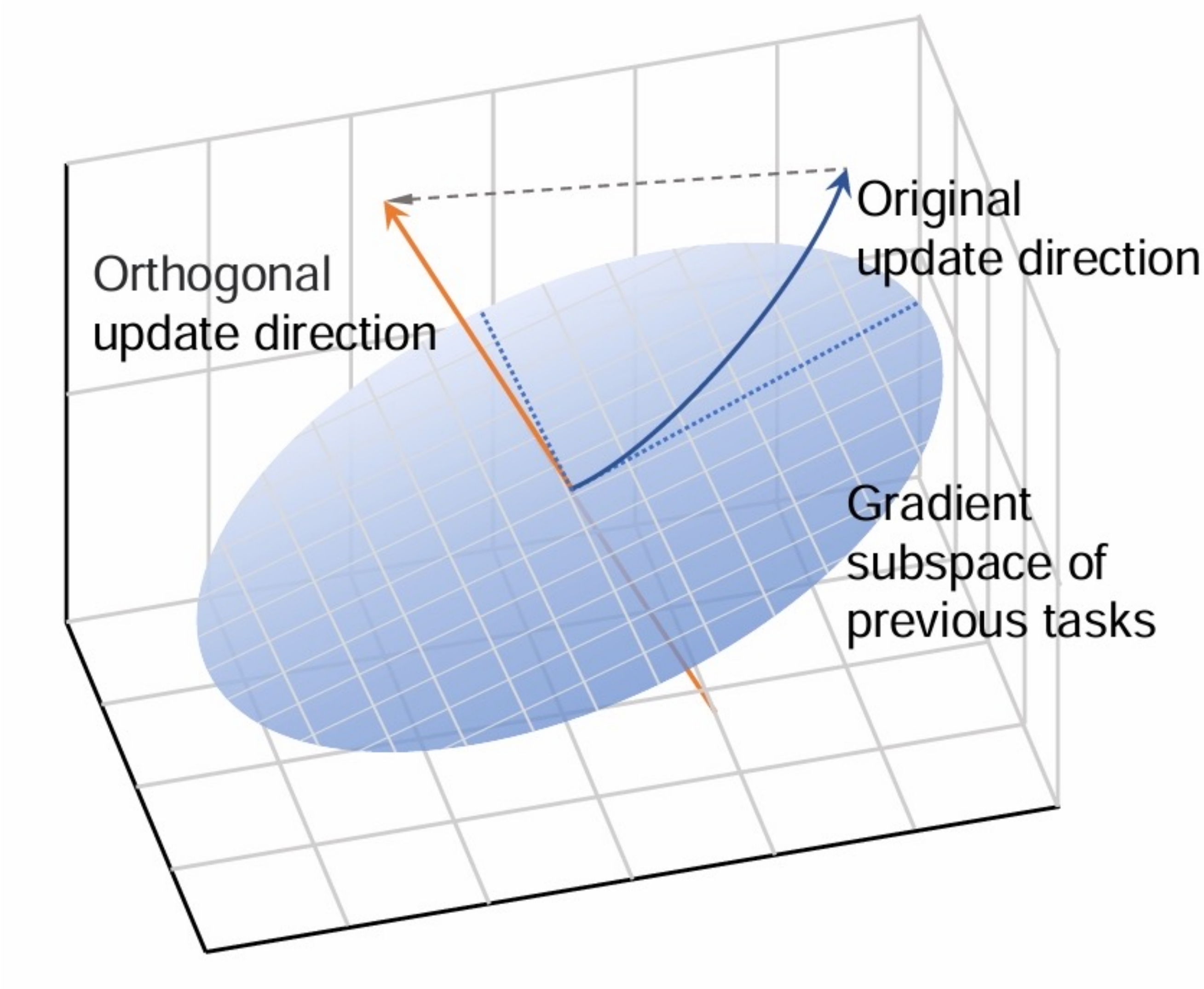
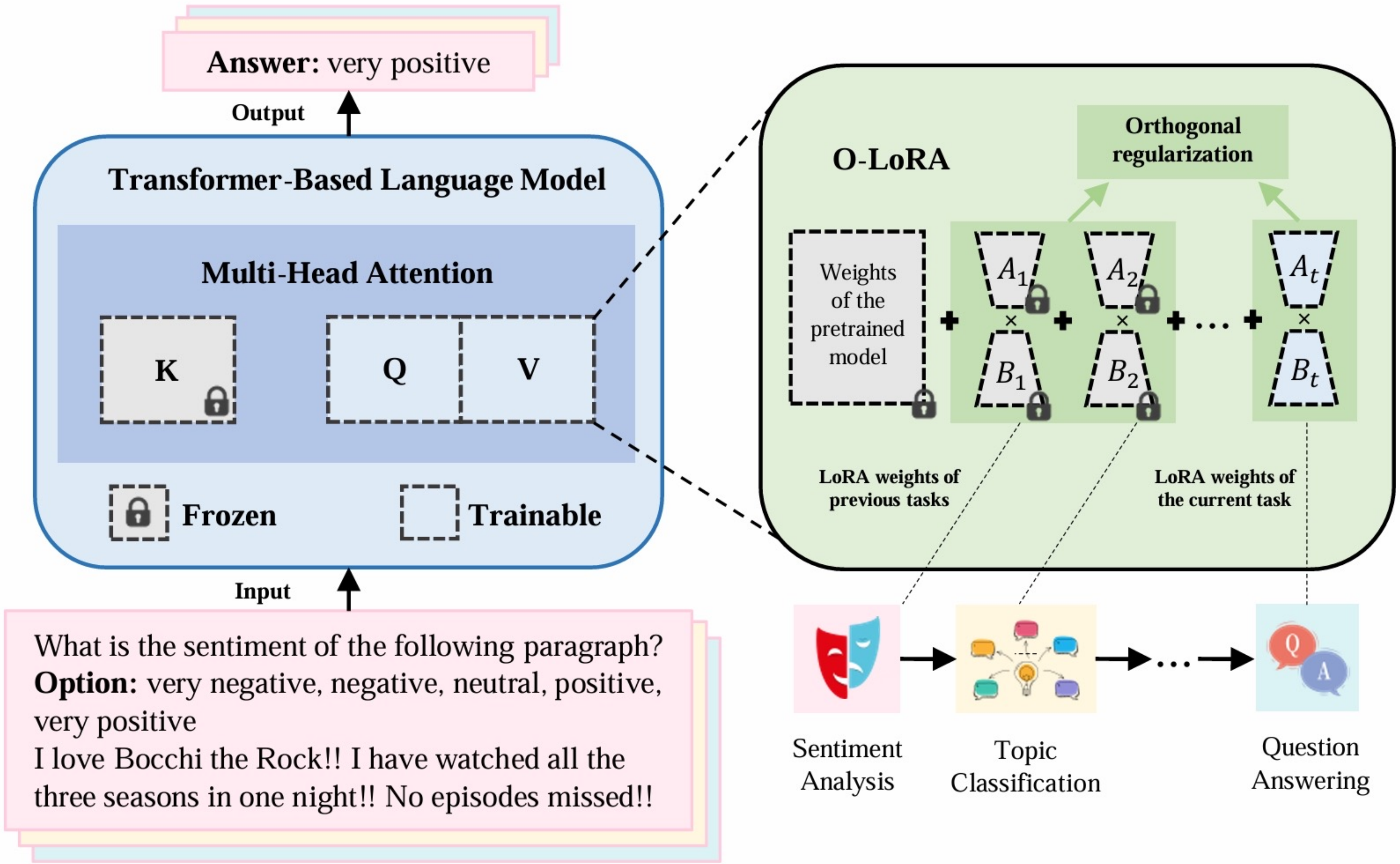


Figure 1: An illustration of how Orthogonal Gradient Descent corrects the directions of the gradients. g is the original gradient computed for task B and $\tilde{g}$ is the projection of g onto the orthogonal space $w.r.t$ the gradient $\nabla f_j(x; w_A^*)$ computed at task A. Moving within this (blue) space allows the model parameters to get closer to the low error (green) region for both tasks.



Optimization direction conflicting



$$\sum_{x,y\in\mathcal{D}_t} \log p_\Theta(y \mid x) + \lambda_1 \sum_{i=1}^{t-1} L_{orth}(A_i, A_t) \quad (7)$$

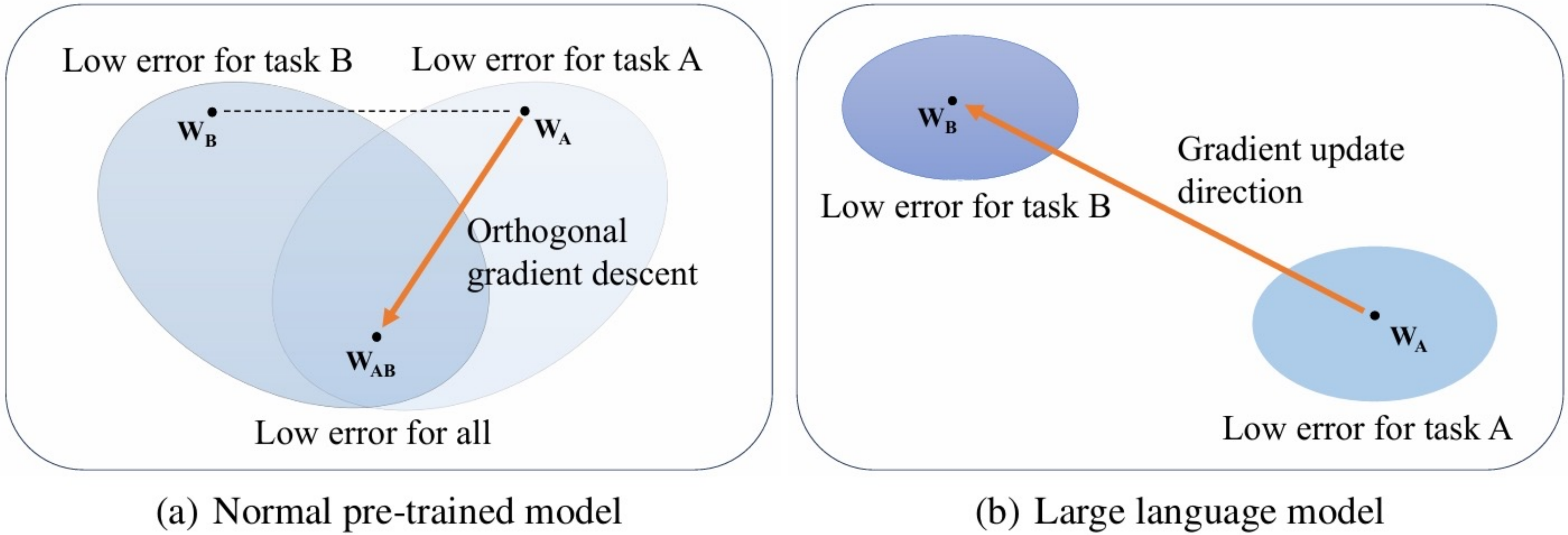
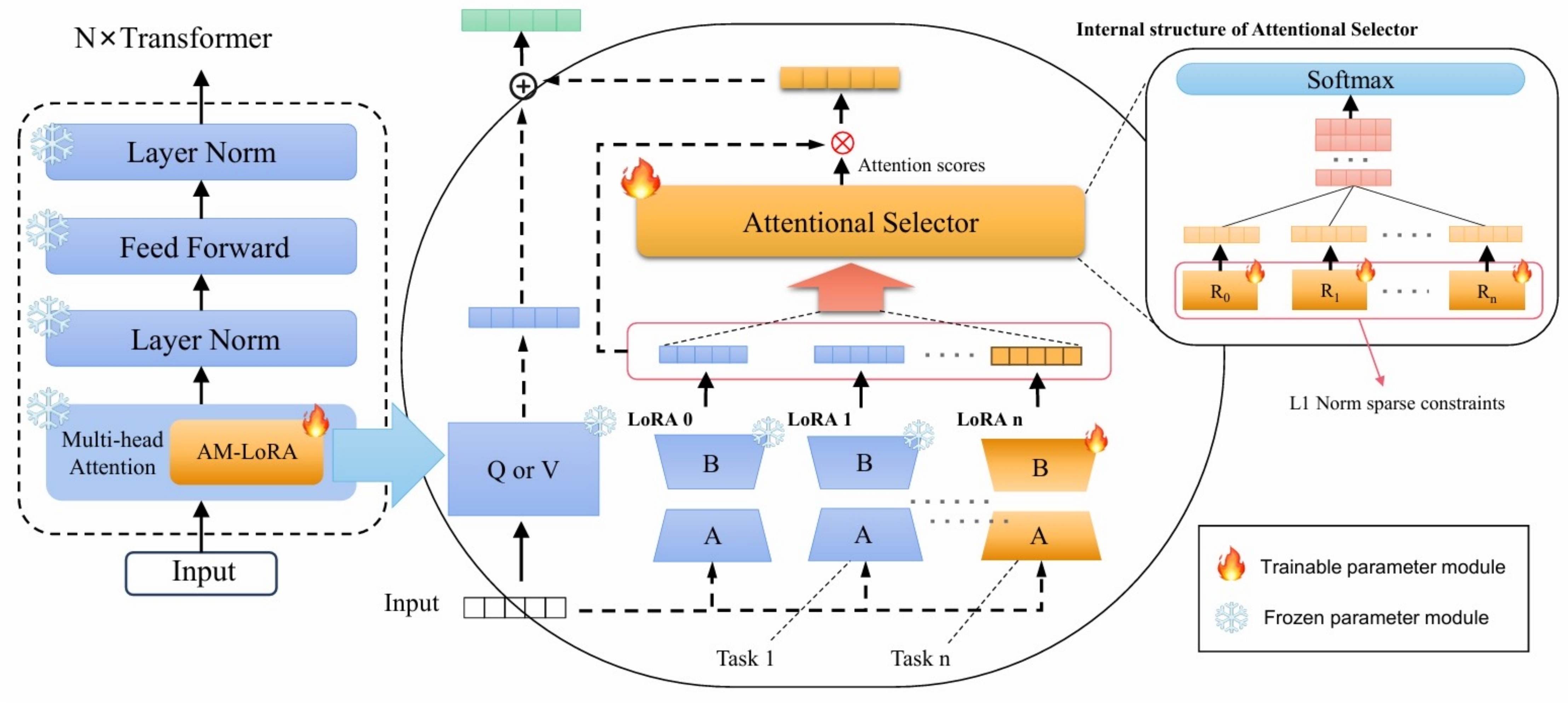


Figure 1: Intuitive demonstration of the decentralized problem of optimal solutions. (a) is the distance relationship diagram between the possible optimal solutions of the two tasks in the normal pre-trained language model. There is a common optimal solution at the intersection of the two. The parameter space of the LLM (b) may be too large, causing the optimal solution areas of the two tasks to be too far apart, so that there is no common optimal solution.



$$g_i = \text{Softmax}(R_i(\Delta w_i x)) = \text{Softmax}(W_{ri}^T(\Delta w_i x)),$$

$$h = W_0 + \sum_{i=0}^{n} g_i \cdot (\Delta w_i x).$$

	Standard CL benchmarks				Large Number of Tasks			
	Order1	**Order2**	**Order3**	**Avg**	**Order4**	**Order5**	**Order6**	**Avg**
SeqFT	18.9	24.9	41.7	28.5	7.4	7.4	7.5	7.4
SinLoRA	44.6	32.7	53.7	43.7	2.3	0.6	1.9	1.6
IncLoRA	66	64.9	68.3	66.4	63.3	58.5	61.7	61.2
Replay	55.2	56.9	61.3	57.8	55	54.6	53.1	54.2
EWC	48.7	47.7	54.5	50.3	45.3	44.5	45.6	45.1
L2P	60.3	61.7	61.1	60.7	57.5	53.8	56.9	56.1
LFPT5	67.6	72.6	**77.9**	72.7	70.4	68.2	69.1	69.2
O-LoRA	75.4	75.7	76.3	75.8	72.3	64.8	71.6	69.6
AM-LoRA	**78.1**	**79.8**	76.2	**78.0**	**72.7**	**73.3**	**71.8**	**72.6**
ProgPrompt	75.2	75	75.1	75.1	78	77.7	77.9	77.9
PerTaskFT	70	70	70	70	78.1	78.1	78.1	78.1
MTL	80	80	80	80	76.5	76.5	76.5	76.5

Table 1: Summary of results on Standard CL benchmarks and Large Number of Tasks benchmarks using T5-large models with AM-LoRA. Report the average accuracy of all tasks after training for the last task. All results were averaged over 3 runs.

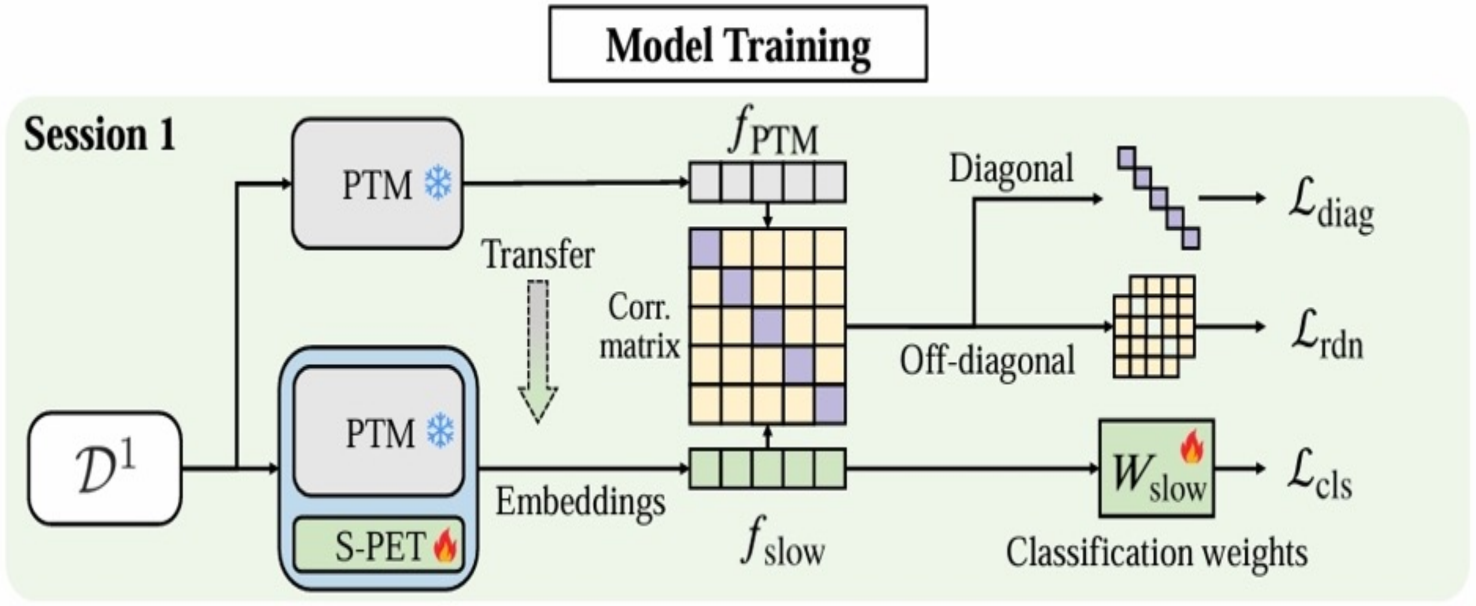
SAFE

Slow learner 细粒度 特征蒸馏

$$\boldsymbol{M}_{i,j} = \frac{1}{N_b}\sum_{k=1}^{N_b}[\phi_{\mathrm{PTM}}(x_k)]_i \cdot [\phi_{\mathrm{slow}}(x_k)]_j,$$

$$\mathcal{L}_{\mathrm{diag}} = \frac{1}{d}\sum_{i=1}^{d}(1-\boldsymbol{M}_{i,i})^2.$$

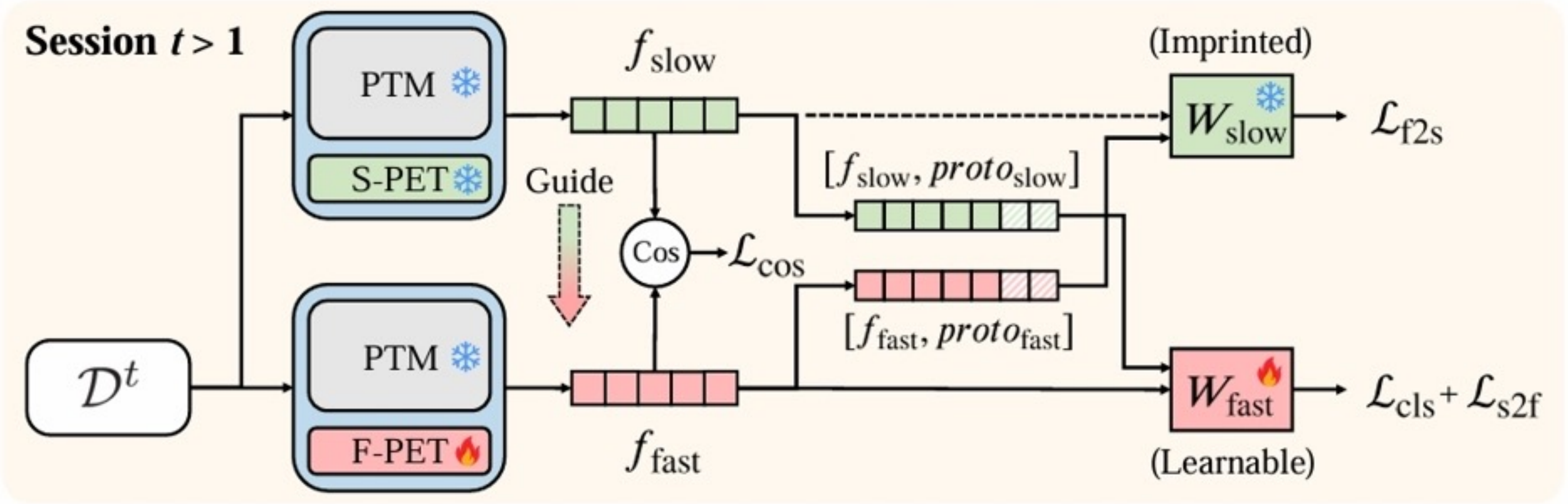
$$\mathcal{L}_{\mathrm{rdn}} = \frac{1}{(d-1)^2}\sum_{i=1}^{d}\sum_{j\neq i}\boldsymbol{M}_{i,j}^2.$$



$$\mathcal{L}_{\mathrm{cls}} = \frac{1}{N_b}\sum_{i=1}^{N_b}\mathrm{CE}(W_{\mathrm{slow}}^{\top}\odot\phi_{\mathrm{slow}}(x_i), y_i),$$

$$\mathcal{L}_{\mathrm{initial}} = \mathcal{L}_{\mathrm{cls}} + \lambda_{\mathrm{diag}}\cdot\mathcal{L}_{\mathrm{diag}} + \lambda_{\mathrm{rdn}}\cdot\mathcal{L}_{\mathrm{rdn}}.$$

Fast learner 细粒度蒸馏 & 校准



$$\mathcal{L}_{\text{cos}} = \frac{1}{N_b}\sum_{i=1}^{N_b}\left(1-\cos(\phi_{\text{slow}}(x_i), \phi_{\text{fast}}(x_i))\right),$$

$$\mathcal{L}_{\text{f2s}} = \frac{1}{N_b}\sum_{i=1}^{N_b}\text{CE}(W_{\text{slow}}^{\top}\odot\phi_{\text{fast}}(x_i), y_i) + \frac{1}{|\mathcal{Y}_{1:t-1}|}\sum_{j=1}^{|\mathcal{Y}_{1:t-1}|}\text{CE}(W_{\text{slow}}^{\top}\odot W_{\text{fast}}^{(j)}, j),$$

$$\mathcal{L}_{\text{follow}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{f2s}} + \mathcal{L}_{\text{s2f}} + \lambda_{\text{cos}}\cdot\mathcal{L}_{\text{cos}},$$

Method	Replay	CIFAR	IN-R	IN-A	CUB	OB	VTAB	Avg
SLCA [49]	w/	91.5	77.0	59.8	84.7	73.1	89.2	79.2
SSIAT [35]		91.4	79.6	62.2	88.8	-	94.5	-
L2P [42]	w/o	84.6	72.5	42.5	65.2	64.7	77.1	67.8
DualPrompt [41]		81.3	71.0	45.4	68.5	65.5	81.2	68.8
CODAPrompt[33]		86.3	75.5	44.5	79.5	68.7	87.4	73.7
ADaM [51]		87.6	72.3	52.6	87.1	74.3	84.3	76.4
EASE [53]		87.8	76.2	55.0	86.8	74.9	93.6	79.1
RanPAC [23]		92.2	78.1	61.8	90.3	79.9	92.6	82.5
SAFE (ours)	w/o	**92.8**	**81.0**	**66.6**	**91.1**	**80.9**	**95.0**	**84.6**

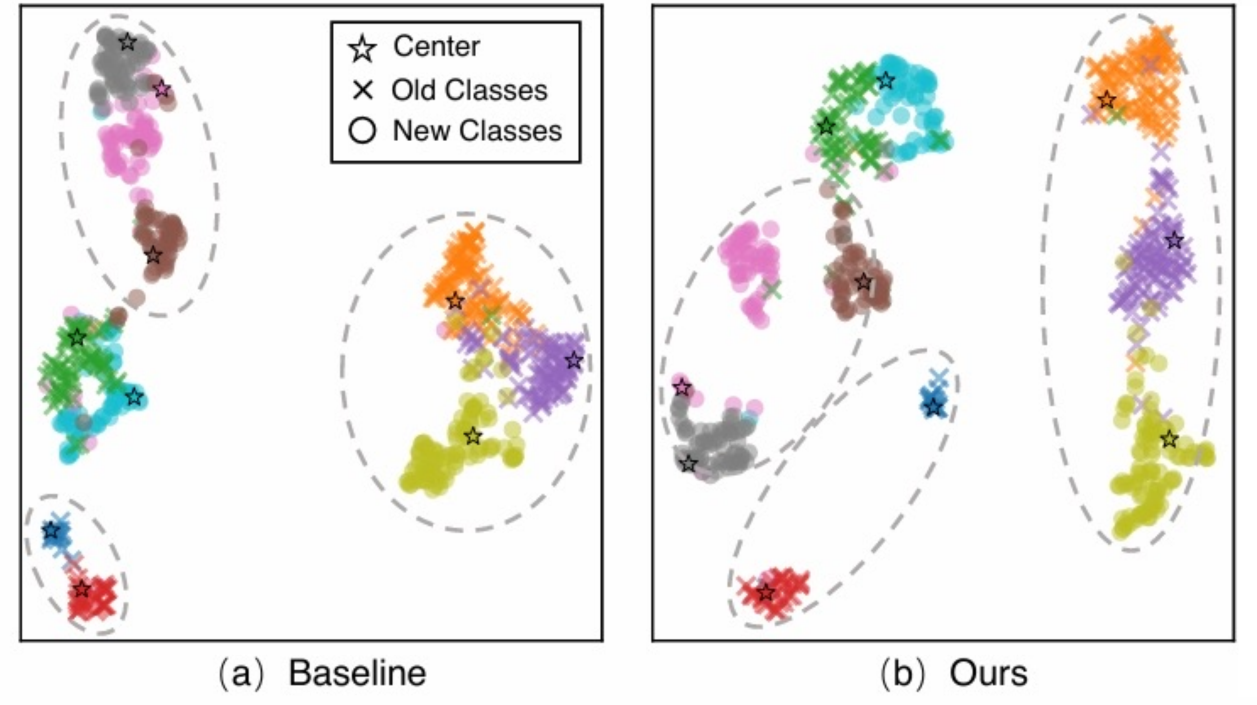


Figure 3: Comparisons with T-SNE visualization

Table 4: Ablation study of slow lear

Method	Final	Average
Baseline	62.21	72.31
Baseline w/ FA	62.81	73.35
Baseline w/ SSA	63.20	73.00
Slow Learner	**65.44**	**74.41**

Table 2: Overall ablation study on IN-A

Method	SL	FL	Final	Averag
Baseline			62.21	72.31
Slow Learner	✓		65.44	74.41
Fast Learner		✓	66.49	74.50
SAFE	✓	✓	**66.56**	**74.71**

1.用reference dataset 蒸馏G-adapter
2.每一个任务均用一个正交lora-adapter，通过DDAS获得的id进行推理