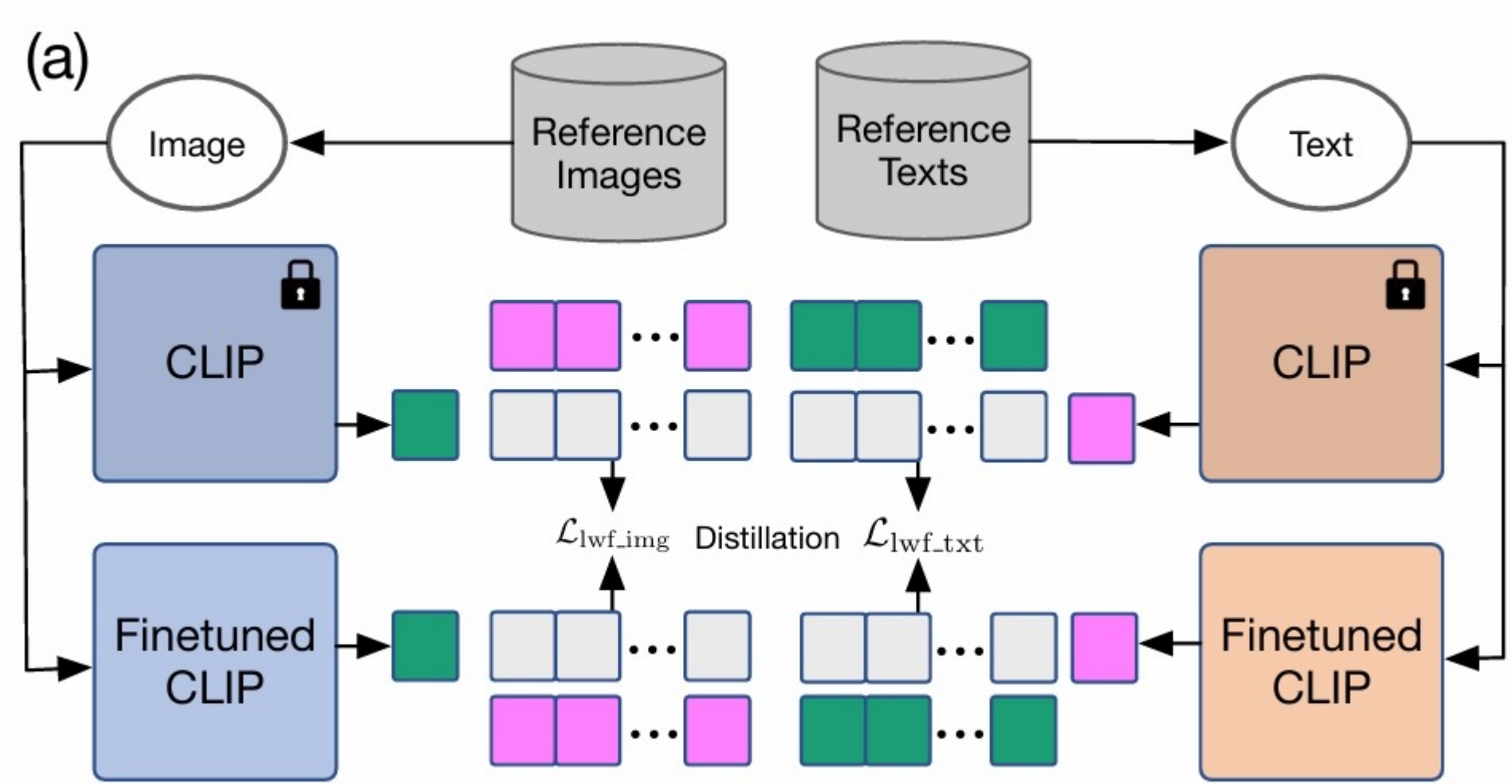
VL IL

Methods	10 steps		20 steps		50 steps	
	Avg	**Last**	**Avg**	**Last**	**Avg**	**Last**
iCaRL (Rebuffi et al., 2017)	65.27	50.74	61.20	43.74	56.08	36.62
UCIR (Hou et al., 2019)	58.66	43.39	58.17	40.63	56.86	37.09
BiC (Wu et al., 2019)	68.80	53.54	66.48	47.02	62.09	41.04
RPSNet (Rajasegaran et al., 2019b)	68.60	57.05	-	-	-	-
WA (Zhao et al., 2020)	69.46	53.78	67.33	47.31	64.32	42.14
PODNet (Douillard et al., 2020)	58.03	41.05	53.97	35.02	51.19	32.99
DER (w/o P) (Yan et al., 2021)	**75.36**	65.22	74.09	62.48	72.41	59.08
DER (Yan et al., 2021)	74.64	64.35	73.98	62.55	72.05	59.76
DyTox (Douillard et al., 2022)	67.33	51.68	67.30	48.45	64.39	43.47
DyTox+ (Douillard et al., 2022)	74.10	62.34	71.62	57.43	68.90	51.09
Continual-CLIP	75.17	**66.72**	**75.95**	**66.72**	**76.49**	**66.72**

Methods	ImageNet100-B0		ImageNet1K		ImageNet100-B50	
	Avg	**Last**	**Avg**	**Last**	**Avg**	**Last**
iCaRL (Rebuffi et al., 2017)	-	-	38.40	22.70	-	-
UCIR (Hou et al., 2019)	-	-	-	-	68.09	57.30
WA (Zhao et al., 2020)	-	-	65.67	55.60	-	-
TPCIL (Tao et al., 2020)	-	-	-	-	74.81	66.91
PODNet (Douillard et al., 2020)	-	-	-	-	74.33	-
Simple-DER (Li et al., 2021b)	-	-	66.63	59.24	-	-
DER (w/o P) (Yan et al., 2021)	77.18	66.70	68.84	60.16	78.20	74.92
DER (Yan et al., 2021)	76.12	66.06	66.73	58.62	77.13	72.06
DyTox (Douillard et al., 2022)	73.96	62.20	-	-	-	-
DyTox+ (Douillard et al., 2022)	77.15	67.70	70.88	60.00	-	-
Continual-CLIP	**85.00**	**75.42**	**75.51**	**67.71**	**79.69**	**75.42**

Preventing Zero-Shot Transfer Degradation in Continual Learning of Vision-Language Models

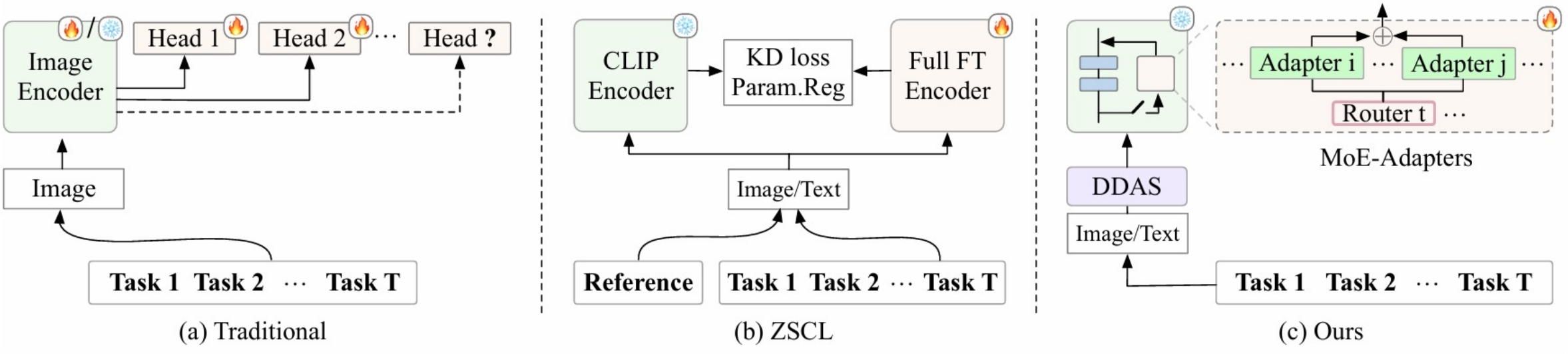


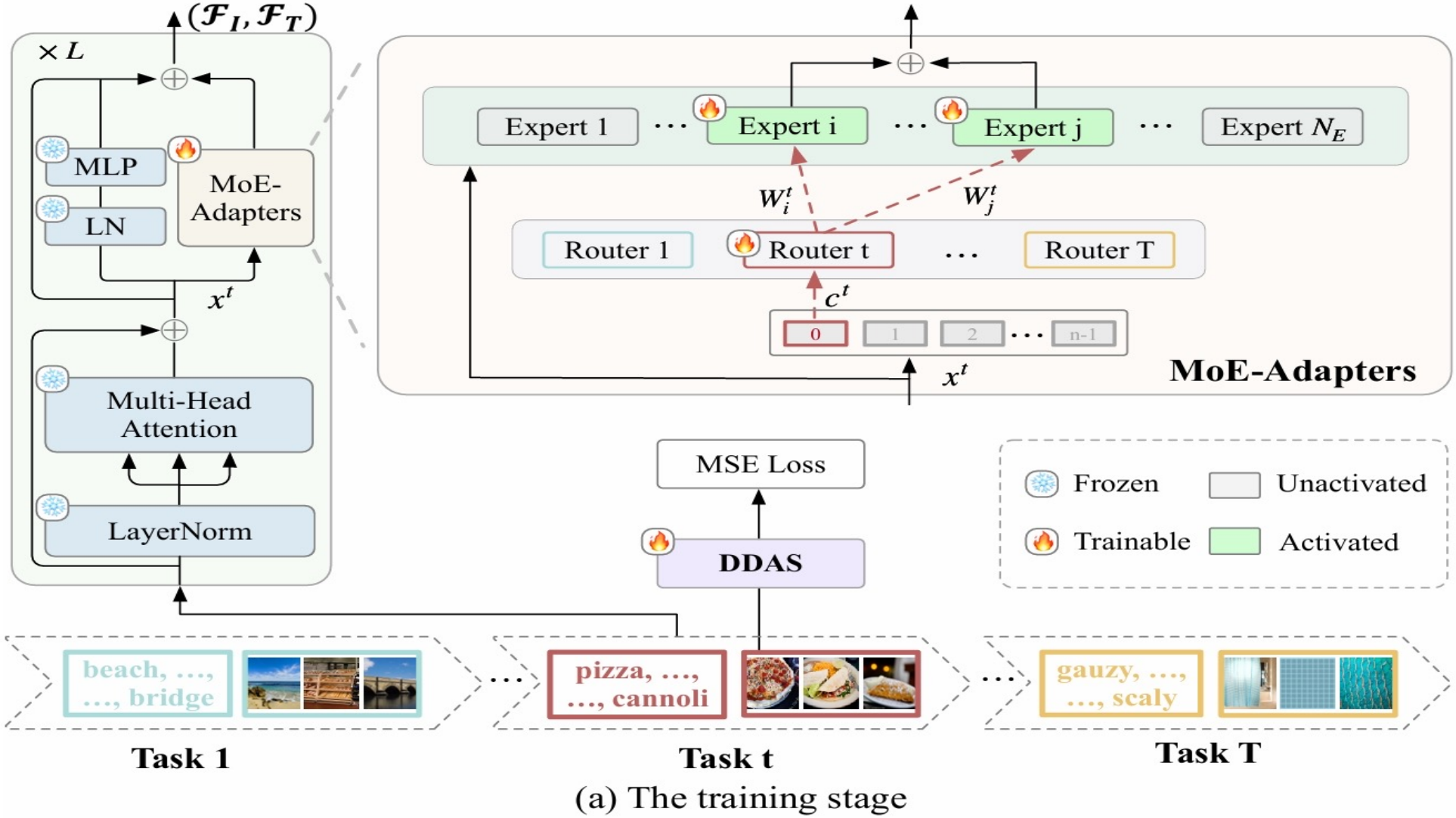
$$\mathcal{L}_{\text{dist\_img}} = \text{CE}(\boldsymbol{p}, \overline{\boldsymbol{p}}) = -\sum_{j=1}^{m} \boldsymbol{p}_j \cdot \log \overline{\boldsymbol{p}}_j,$$

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \lambda \cdot (\mathcal{L}_{\text{lwf\_img}} + \mathcal{L}_{\text{lwf\_txt}}).$$

$$\hat{\theta}_t = \begin{cases} \theta_0 & t = 0 \\ \frac{1}{t+1}\theta_t + \frac{t}{t+1} \cdot \hat{\theta}_{t-1} & \text{every I iterations} \end{cases}.$$

Preventing Zero-Shot Transfer Degradation in Continual Learning of Vision-Language Models

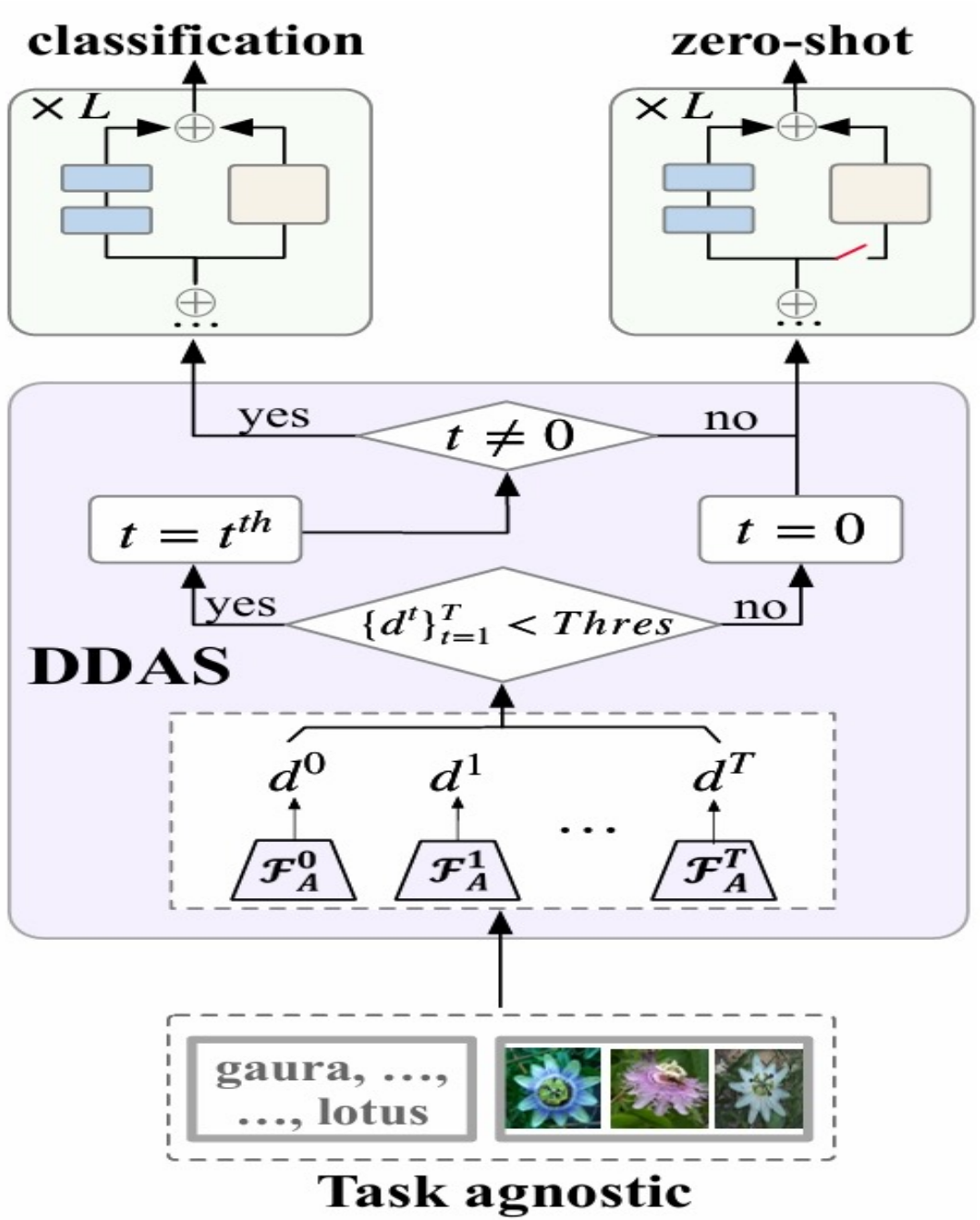




(a) The training stage

$$\boldsymbol{y}^t = \sum_{i=1}^{N_E} W_i^t \mathcal{E}_i(\boldsymbol{x}^t),$$

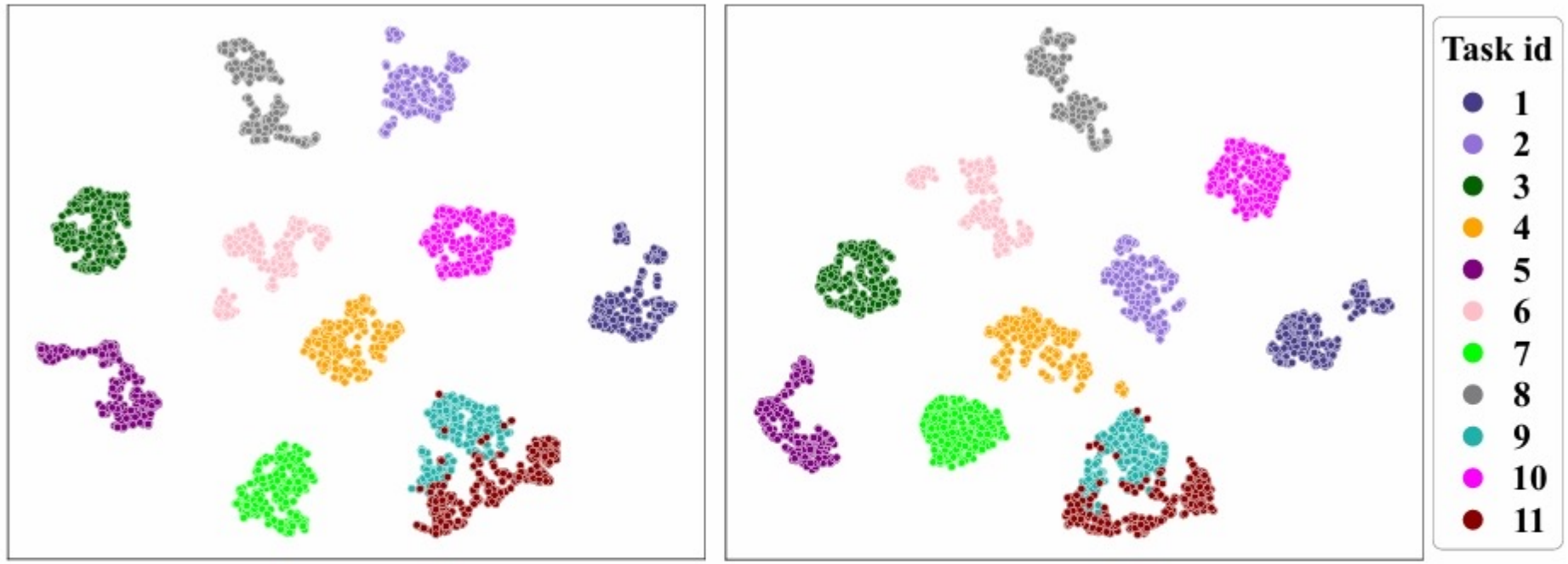
$$W^t = Softmax(Topk(\mathcal{R}^t(\boldsymbol{c}^t))),$$



(b) The inference stage

$$d^t = ||\boldsymbol{f}_i^t - \boldsymbol{f}_o^t||^2,$$

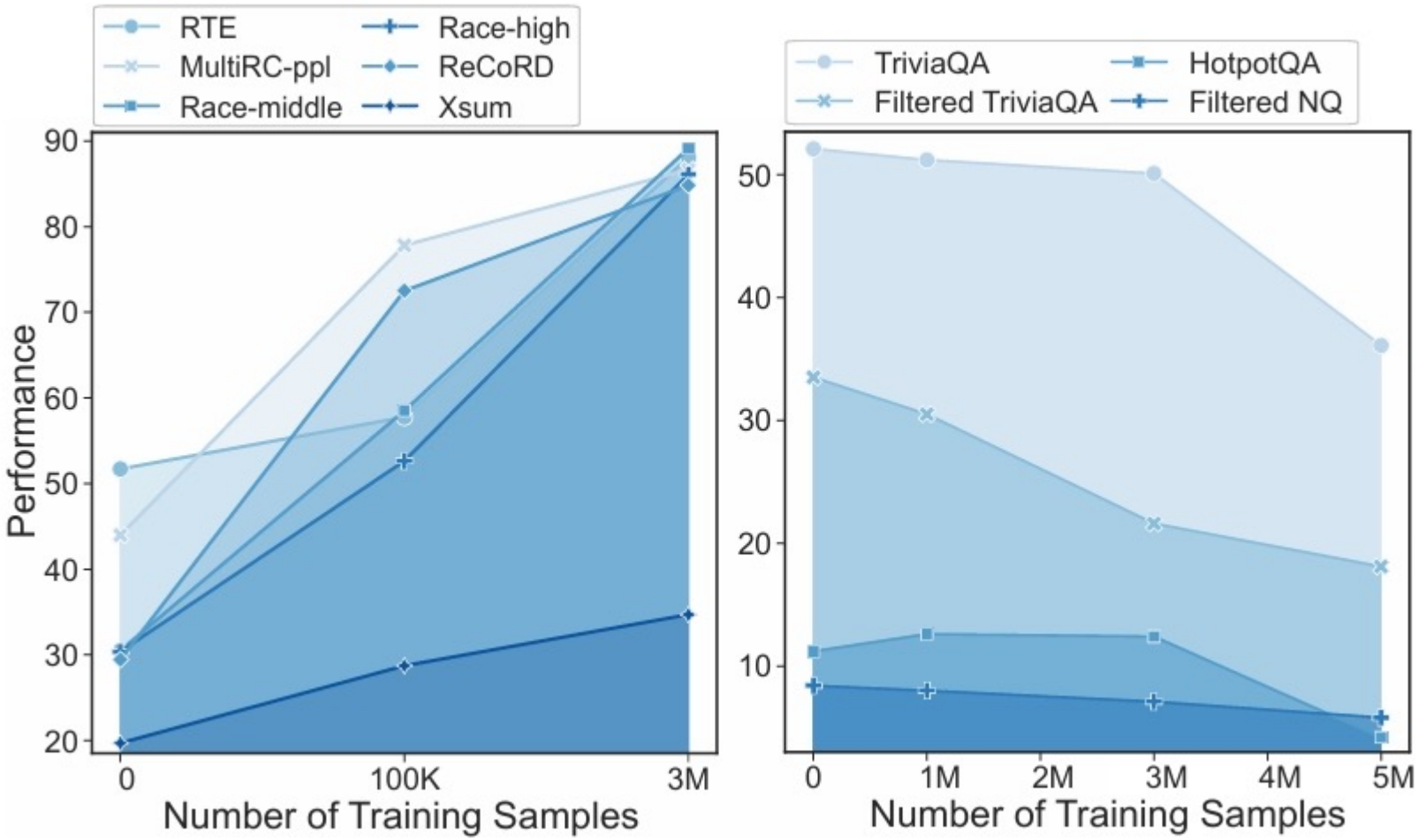
	Method	Aircraft [49]	Caltech101 [21]	CIFAR100 [38]	DTD [9]	EuroSAT [25]	Flowers [54]	Food [4]	MNIST [13]	OxfordPet [58]	Cars [37]	SUN397 [69]	*Average*
CLIP	Zero-shot	24.3	88.4	68.2	44.6	54.9	71.0	88.5	59.4	89.0	64.7	65.2	65.3
	5-shot Full Fine-tune	30.6	93.5	76.8	65.1	91.7	92.9	83.3	96.6	84.9	65.4	71.3	77.5
	5-shot Fine-tune Adapter	29.7	90.0	75.3	63.9	81.1	94.2	87.8	90.4	89.0	68.2	72.5	76.6
Transfer	Continual-FT		72.8	53.0	36.4	35.4	43.3	68.4	47.4	72.6	30.0	52.7	51.2
	LwF [42]		72.1	49.2	35.9	44.5	41.1	66.6	50.5	69.0	19.0	51.7	50.0
	LwF-VR [15]		82.2	62.5	40.1	40.1	56.3	80.0	60.9	77.6	40.5	60.8	60.1
	WiSE-FT [67]		77.6	60.0	41.3	39.4	53.0	76.6	58.1	75.5	37.3	58.2	57.7
	ZSCL [78]		84.0	68.1	**44.8**	46.8	63.6	84.9	61.4	81.4	55.5	62.2	65.3
	Ours		**87.9**	**68.2**	44.1	**48.1**	**64.7**	**88.8**	**69.0**	**89.1**	**64.5**	**65.1**	**68.9**(+3.6)
Average	Continual-FT	28.1	86.4	59.1	52.8	55.8	62.0	70.2	64.7	75.5	35.0	54.0	58.5
	LwF [42]	23.5	77.4	43.5	41.7	43.5	52.2	54.6	63.4	68.0	21.3	52.6	49.2
	LwF-VR [15]	24.9	89.1	64.2	53.4	54.3	70.8	79.2	66.5	79.2	44.1	61.6	62.5
	WiSE-FT [67]	**32.0**	87.7	61.0	55.8	68.1	69.3	76.8	71.5	77.6	42.0	59.3	63.7
	ZSCL [78]	28.2	88.6	66.5	53.5	56.3	73.4	83.1	56.4	82.4	57.5	62.9	64.4
	Ours	30.0	**89.6**	**73.9**	**58.7**	**69.3**	**79.3**	**88.1**	**76.5**	**89.1**	**65.3**	**65.8**	**71.4**(+7.0)
Last	Continual-FT	27.8	86.9	60.1	58.4	56.6	75.7	73.8	93.1	82.5	57.0	66.8	67.1
	LwF [42]	22.1	58.2	17.9	32.1	28.1	66.7	46.0	84.3	64.1	31.5	60.1	46.5
	LwF-VR [15]	22.9	**89.8**	59.3	57.1	57.6	79.2	78.3	77.7	83.6	60.1	69.8	66.9
	WiSE-FT [67]	**30.8**	88.9	59.6	60.3	80.9	81.7	77.1	**94.9**	83.2	62.8	70.0	71.9
	ZSCL [78]	26.8	88.5	63.7	55.7	60.2	82.1	82.6	58.6	85.9	66.7	70.4	67.4
	Ours	30.1	89.3	**74.9**	**64.0**	**82.3**	**89.4**	**87.1**	89.0	**89.1**	**69.5**	**72.5**	**76.1**(+4.2)

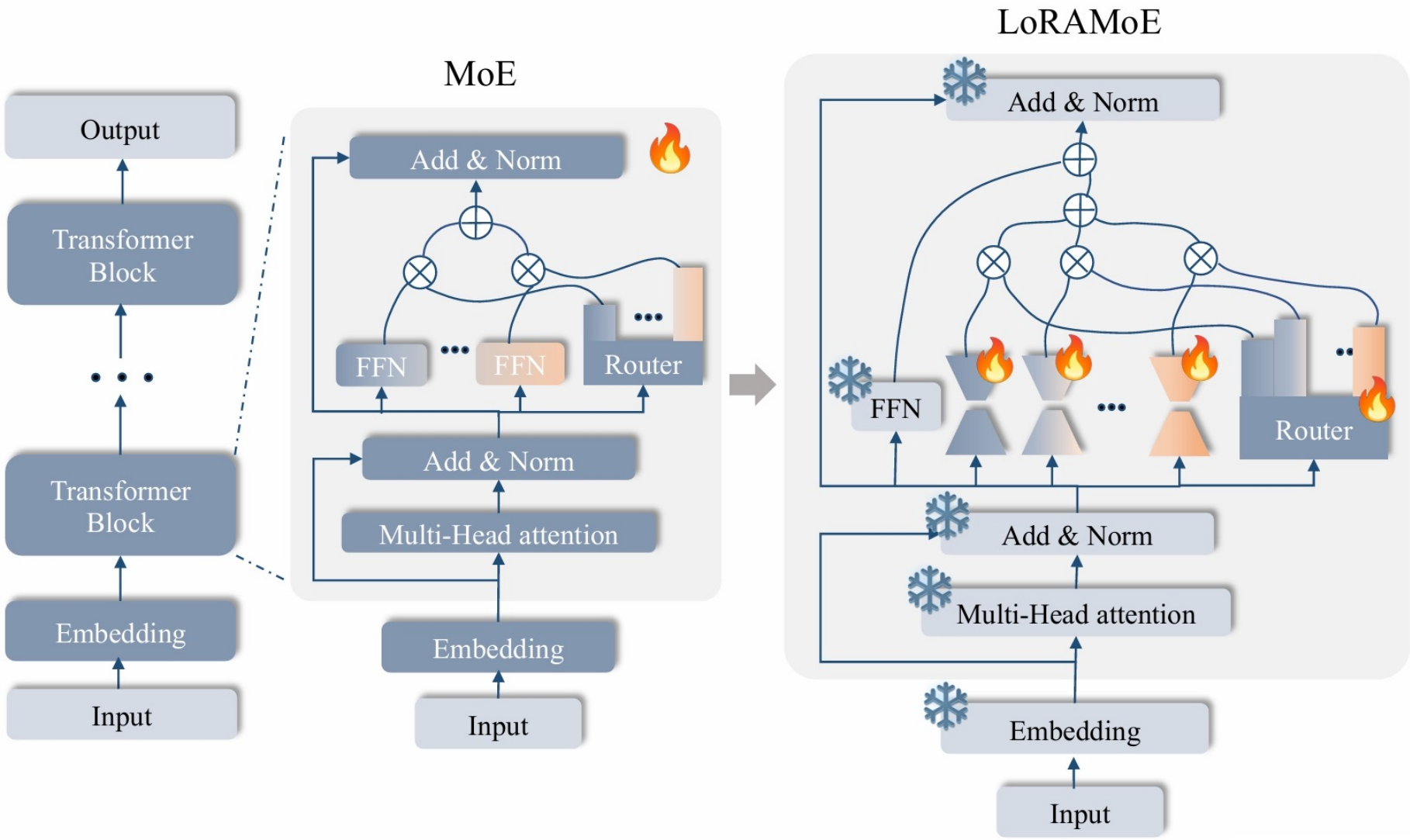


(a) Full-shot **(b) Few-shot**

Figure 5. t-SNE on DDAS's output of each task on full-shot and few-shot MTIL. The corresponding task names from $id = 1 - 11$ are matches with the datasets listed from left to right in Table 1.

LoraMOE


RTE
MultiRC-ppl
Race-middle
Race-high
ReCoRD
Xsum
Performance
90
80
70
60
50
40
30
20
0
100K
3M
Number of Training Samples
TriviaQA
Filtered TriviaQA
HotpotQA
Filtered NQ
50
40
30
20
10
0
1M
2M
3M
4M
5M
Number of Training Samples



$$o = W_0 x + \Delta W x = W_0 x + \sum_{i=1}^{N} G(x)_i E_i(x) \quad (3)$$

$$\Delta W_E = BA$$

$$o = W_0 x + \frac{\alpha}{r} \sum_{i=1}^{N} \omega_i \cdot B_i A_i x$$

$$G(\cdot) = Softmax(xW_g)$$

1.Experts 数量　10 → 6
2.路由权重分布
3.DDAS具体精度