Table 6: Comparison of different MTL methods on CIFAR100 and VOC07+12 datasets. The 1^{st}/2^{nd} best results are indicated in red/blue.

Method	Before evolution (%) ↑			After evolution (%) ↑		
	CLS	DET	AVG	CLS	DET	AVG
Expert(CLS)+MAD	70.68±0.00	0.00±0.00	35.34±0.00	70.52±0.17	82.58±0.13	**76.55±0.15**
Expert(CLS)+Pix2SeqV2	70.68±0.00	0.00±0.00	35.34±0.00	70.24±0.15	80.14±0.09	75.19±0.12
Expert(CLS)+Frozen	70.68±0.00	0.00±0.00	35.34±0.00	**70.68±0.00**	59.92±0.06	65.30±0.03
Expert(DET)+MAD	70.68±0.00	0.00±0.00	35.34±0.00	70.52±0.17	82.58±0.13	76.55±0.15
Expert(DET)+Pix2SeqV2	70.68±0.00	0.00±0.00	35.34±0.00	70.24±0.15	80.14±0.09	75.19±0.12
Expert(DET)+Frozen	0.47±0.00	83.89±0.00	42.18±0.00	43.05±0.08	**83.89±0.00**	63.47±0.04
DISC	70.68±0.00	0.00±0.00	35.34±0.00	**72.28±0.08**$_{(+1.60)}$	**84.92±0.04**$_{(+1.03)}$	**78.60±0.06**$_{(+2.05)}$

Table 7: The relationship from human society to machine society to method design.

Human Society	**Machine Society**	**Method Design**
Social hierarchy	Hierarchical organizational structures	Layer-wise hierarchical structures
Sequential progression	Progressive interaction modes	Dynamic hierarchical collaboration
Cultural learning	Strong-guided communication mechanisms	Dynamic selective collaboration

Table 8: Comparison of different backbones, training strategies and tasks on CIFAR100 and VOC07+12 datasets. The 1^{st}/2^{nd} best results are indicated in red/blue.

Method	CLS (%) ↑	DET (%) ↑	SEG (%) ↑	AVG (%) ↑
ResNet-50 (Scratch)	61.15±0.08	40.11±0.06	50.83±0.12	50.69±0.09
ResNet-50 (ImageNet)	64.02±0.03	67.74±0.09	55.63±0.04	62.46±0.05
ResNet-101 (ImageNet)	**64.73±0.09**	**69.66±0.05**	**56.29±0.06**	**63.56±0.07**
ResNet-152 (ImageNet)	**64.97±0.06**$_{(+0.24)}$	**71.18±0.07**$_{(+1.52)}$	**57.68±0.03**$_{(+1.39)}$	**64.61±0.05**$_{(+1.05)}$

Table 9: Comparison of different knowledge acquisition methods on CIFAR100 and VOC07+12 datasets. The 1^{st}/2^{nd} best results are indicated in red/blue.

Method	CLS (%) ↑	DET (%) ↑	AVG (%) ↑
Data (Direct Experience)	69.84±0.07	82.96±0.11	76.40±0.09
Data-augmentation (Direct Experience)	70.68±0.09	83.89±0.07	77.29±0.08
General-Model (Indirect Experience)	70.95±0.07	83.91±0.08	77.43±0.08
Specialist-Model (Indirect Experience)	**71.13±0.15**	**84.01±0.12**	**77.57±0.14**
DISC (Direct and Indirect Experience)	**72.28±0.08**$_{(+1.15)}$	**84.92±0.04**$_{(+0.91)}$	**78.60±0.06**$_{(+1.03)}$

Table 10: Comparison of performance across different tasks on CIFAR10, Food-101 and WIDER FACE(E, M, H) datasets. The 1^{st}/2^{nd} best results are indicated in red/blue.

Method	CLS (CIFAR10) ↑	CLS (Food-101) ↑	DET (WIDER FACE (E)) ↑	DET (WIDER FACE (M)) ↑	DET (WIDER FACE (H)) ↑	AVG (%) ↑
LSKD(CLS)+MTL	89.03±0.05	77.16±0.17	66.28±0.11	66.53±0.16	50.22±0.14	69.84±0.13
LSKD(DET)+MTL	68.26±0.09	56.02±0.15	88.04±0.16	87.15±0.12	71.18±0.11	74.13±0.13
CrossKD(CLS)+MTL	**89.72±0.08**	**78.01±0.17**	68.15±0.12	67.31±0.17	51.25±0.14	70.89±0.14
CrossKD(DET)+MTL	69.04±0.05	57.28±0.12	**88.92±0.16**	**87.99±0.12**	**71.78±0.11**	**75.00±0.11**
PPAL(CLS)+MTL	89.15±0.09	77.35±0.13	67.24±0.16	67.05±0.11	50.72±0.17	70.30±0.13
PPAL(DET)+MTL	68.62±0.07	56.63±0.17	88.31±0.12	87.63±0.15	71.53±0.16	74.54±0.13
DISC	**90.98±0.05**$_{(+1.26)}$	**79.25±0.12**$_{(+1.24)}$	**90.82±0.15**$_{(+1.90)}$	**89.35±0.13**$_{(+1.36)}$	**73.02±0.16**$_{(+1.24)}$	**84.68±0.12**$_{(+9.68)}$

Table 11: Comparison of computational cost.

Method	**FLOPs (G)**	**Params (M)**
LSKD(CLS)	0.14	0.47
LSKD(DET)	52.96	22.59
CrossKD(CLS)	0.16	26.91
CrossKD(DET)	83.32	32.22
PPAL(CLS)	0.17	29.76
PPAL(DET)	85.05	36.45
MAD(CLS)	0.26	90.62
MAD(DET)	194.39	107.85
Pix2SeqV2(CLS)	0.29	116.89
Pix2SeqV2(DET)	204.81	132.62
DISC(CLS)	0.24	67.69
DISC(DET)	152.47	85.42

Proof. (Dynamic over static) Let $D_{\text{train}} = \{x_i, y_i\}_{i=1}^{N}$ be a training dataset of N samples, and $\hat{\mathbb{E}}(f^m)$ be the empirical error of the m-th model f^m on D_{train}. For any hypothesis $f \in \mathcal{H}$ (i.e., $\mathcal{H} : \mathcal{X} \to \{-1, 1\}$), with probability at least $1 - \delta$, the generalization error is bounded by:

$$\text{GError}(f) \leq \underbrace{\sum_{m=1}^{M} \mathbb{E}(w^m)\hat{\mathbb{E}}(f^m)}_{\text{Term-L (empirical loss)}} + \underbrace{\sum_{m=1}^{M} \mathbb{E}(w^m)\mathfrak{R}_m(f^m)}_{\text{Term-C (complexity)}} + \underbrace{\sum_{m=1}^{M} Cov(w^m, \ell^m)}_{\text{Term-Cov (covariance)}} + M\sqrt{\frac{\ln(1/\delta)}{2N}}, \tag{1}$$

where $\mathbb{E}(w^m)$ is the expected collaboration weight, $\mathfrak{R}_m(f^m)$ is the Rademacher complexity of model f^m, and $Cov(w^m, \ell^m)$ is the covariance between the weight and the loss.

In *static collaboration*, the weights w^m_{static} are constant, hence

$$Cov(w^m_{\text{static}}, \ell^m) = 0. \tag{2}$$

In *dynamic collaboration*, the collaboration weight w^m_{dynamic} increases as the model loss ℓ^m decreases. Thus,

$$Cov(w^m_{\text{dynamic}}, \ell^m) < 0, \tag{3}$$

which effectively reduces the Term-Cov and thereby lowers the generalization bound.

Based on the principle of convexity, it can be concluded that:

$$\sum \mathbb{E}(w^m_{\text{dynamic}})\,\hat{\mathbb{E}}(f^m) = \sum w^m_{\text{static}}\,\hat{\mathbb{E}}(f^m). \tag{4}$$

$$\sum \mathbb{E}(w^m_{\text{dynamic}})\,\mathfrak{R}_m(f^m) \leq \sum w^m_{\text{static}}\,\mathfrak{R}_m(f^m). \tag{5}$$

Since the confidence term $M\sqrt{\frac{\ln(1/\delta)}{2N}}$ is independent of the collaboration strategy, it remains the same. Therefore, suppose the hypothesis space is $\mathcal{H}: \mathcal{X} \to \{-1, 1\}$. Then for any $f_{\text{dynamic}}, f_{\text{static}} \in \mathcal{H}$, and for $1 > \delta > 0$, it holds that

$$\mathcal{O}(\text{GError}_{\text{dynamic}}) \leq \mathcal{O}(\text{GError}_{\text{static}}), \tag{6}$$

proving that dynamic collaboration yields a tighter generalization error bound than static collaboration.

□

Proof. (Dynamic weights depend on both itself and the collaborators) Given the dynamic collaboration formula delineated in $f(x) = \sum_{m=1}^{|\mathcal{M}|} \omega^m \cdot f^m(x^m)$, consider ℓ to be the convex logistic loss function. The softmax function is utilized to normalize $\omega^m = \frac{e^{w^m}}{\sum_{j=1}^{|\mathcal{M}|} e^{w^j}}$:.

Considering the property of convex function, we have:

$$\ell(f(x), y) = \ell\left(\sum_{m=1}^{|\mathcal{M}|} \omega^m f^m(x^m), y\right) \leq \sum_{m=1}^{|\mathcal{M}|} \omega^m \ell(f^m(x^m), y). \tag{7}$$

When computing the expectation of Equation (7) and leveraging the properties of expectation, the subsequent equation is satisfied. To simplify notation, $\ell(f(m), y)$ can be denoted as ℓ^m and D is an unknown dataset:

$$\begin{aligned}
GE(f) &= \mathbb{E}_{(x,y)\sim\mathcal{D}} \ell(f(x), y) \\
&\leq \sum_{m=1}^{|\mathcal{M}|} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\omega^m \ell^m\right] \\
&= \frac{1}{|\mathcal{M}|}\left(|\mathcal{M}| \cdot \sum_{m=1}^{|\mathcal{M}|} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\omega^m \ell^m\right]\right) \\
&= \frac{1}{|\mathcal{M}|} \cdot \sum_{m=1}^{|\mathcal{M}|}\left(\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\omega^m \ell^m\right] + (|\mathcal{M}|-1)\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\left(1 - \sum_{j\neq m} \omega^j\right)\ell^m\right]\right) \\
&= \frac{1}{|\mathcal{M}|} \cdot \sum_{m=1}^{|\mathcal{M}|}\Bigg(\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\omega^m\right] \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell^m\right] + Cov(\omega^m, \ell^m) \\
&\quad - (|\mathcal{M}|-1) \cdot \sum_{j\neq m} \left(\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\omega^j\right] \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell^m\right] + Cov(\omega^j, \ell^m) - \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell^m\right]\right)\Bigg) \\
&= \frac{1}{|\mathcal{M}|} \cdot \sum_{m=1}^{|\mathcal{M}|}\Bigg[\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell^m\right]\left(\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\omega^m\right] + (|\mathcal{M}|-1) \cdot \left[1 - \sum_{j\neq m} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\omega^j\right]\right]\right) \\
&\quad + Cov(\omega^m, \ell^m) - (|\mathcal{M}|-1)\sum_{j\neq m} Cov(\omega^j, \ell^m)\Bigg] \\
&= \frac{1}{|\mathcal{M}|} \cdot \sum_{m=1}^{|\mathcal{M}|}\left[|\mathcal{M}| \cdot \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell^m\right] \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\omega^m\right] + Cov(\omega^m, \ell^m) - (|\mathcal{M}|-1)\sum_{j\neq m} Cov(\omega^j, \ell^m)\right] \\
&= \sum_{m=1}^{|\mathcal{M}|}\left(\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell^m\right] \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\omega^m\right] + \frac{1}{|\mathcal{M}|} \cdot \left[Cov(\omega^m, \ell^m) - (|\mathcal{M}|-1)\sum_{j\neq m} Cov(\omega^m, \ell^j)\right]\right) \\
&\leq \sum_{m=1}^{|\mathcal{M}|}\left(\mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell^m\right] + \frac{1}{|\mathcal{M}|} Cov(\omega^m, \ell^m) - \frac{|\mathcal{M}|-1}{|\mathcal{M}|}\sum_{j\neq m} Cov(\omega^m, \ell^j)\right)
\end{aligned} \tag{8}$$

To simplify Equation (8), we invoke Rademacher complexity theory, which establishes that with a confidence level of $1-\Delta$ where $0<\Delta<1$, the following holds:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell^m] \leq \hat{e}rr[f^m] + R_N(\mathcal{H}) + \sqrt{\frac{\ln(1/\Delta)}{2N}}. \tag{9}$$

In this context, $\hat{e}rr(f^m)$ represents the empirical error of the unimodal function f^m, and $\mathcal{H}$ denotes the hypothesis set, defined as $\mathcal{H}: X \to \{-1,+1\}$, which includes f as a member. The Rademacher complexity is denoted by $R_N(\mathcal{H})$. Consequently, we assert that with a confidence level of $1-\Delta$, where $0<\Delta<1$, the following relationship is upheld:

$$\begin{aligned} GE(f) \leq\ & |\mathcal{M}|\left(\mathcal{R}_N(\mathcal{H}) + \sqrt{\frac{\ln(1/\Delta)}{2N}}\right) + \sum_{m=1}^{|\mathcal{M}|} \hat{er}(f^m) \\ & + \sum_{m=1}^{|\mathcal{M}|}\left[\frac{1}{|\mathcal{M}|}\underbrace{Cov(\omega^m,\ell^m)}_{\textit{Negative self-correlation in DSC}} - \frac{|\mathcal{M}|-1}{|\mathcal{M}|}\sum_{j\neq m}\underbrace{Cov(\omega^m,\ell^j)}_{\textit{Positive collaborative correlation in DSC}}\right], \end{aligned} \tag{10}$$

proving that dynamic weights depend on both itself and the collaborator, and that dynamic collaboration leads to a tighter generalization error bound.

□