Table 1: Comparison of Different MTL Methods

Method	**Before evolution (%) ↑**			**After evolution (%) ↑**		
	CLS	**DET**	**AVG**	**CLS**	**DET**	**AVG**
Expert(CLS)+MAD						
Expert(CLS)+VisionLLM						
Expert(DET)+MAD						
Expert(DET)+VisionLLM						
DISC						

Table 2: The Relationship from Human Society to Machine Society to Method Design

Human Society	**Machine Society**	**Method Design**
Social Hierarchy [5]	Hierarchical organizational structures	Layer-wise Hierarchical structures
Sequential Progression [6]	Progressive interaction modes	Dynamic Hierarchical Collaboration
Cultural Learning [4]	Strong-guided communication mechanisms	Dynamic Selective Collaboration

Table 3: Comparison of Different Training Methods

Method	**CLS**	**DET**	**SEG**	**AVG**
ResNet-50 (Scratch)				
ResNet-50 (ImageNet)				
ResNet-50 (GV-D)				
MetaNet-B4 (GV-D)				

Table 4: Comparison of Different Knowledge Acquisition Methods

Method	**CLS**	**DET**	**AVG**
Data (Direct Experience)	69.84	82.96	
Data-augmentation (Direct Experience)	70.68	83.89	
General-Model (Indirect Experience)	70.95	83.91	
Specialist-Model (Indirect Experience)	71.13	84.01	
DISC (Direct and Indirect Experience)	72.28	84.92	

Table 5: Comparison of Different Training Methods on Multiple Datasets

Method	**CLS (CIFAR100)**	**CLS (Food-101)**	**CLS (Caltech-101)**	**DET (VOC07+12)**	**DET (WIDER FACE)**
LSKD(CLS)+MTL					
LSKD(DET)+MTL					
CrossKD(CLS)+MTL					
CrossKD(DET)+MTL					
PPAL(CLS)+MTL					
PPAL(DET)+MTL					
DISC					

Table 6: Comparison of Computational Cost (FLOPs and Params) Across Knowledge Distillation and MTL Methods

Method	**FLOPs**	**Params**
LSKD(CLS)		
LSKD(DET)		
CrossKD(CLS)		
CrossKD(DET)		
PPAL(CLS)		
PPAL(DET)		
MAD(CLS)		
MAD(DET)		
VisionLLM(CLS)		
VisionLLM(DET)		
DISC(CLS)		
DISC(DET)		

We theoretically justify that dynamic fusion outperforms static fusion via a tighter generalization bound.

Proof 1. (Superiority of Dynamic Fusion)

Let $D_{\text{train}} = \{x_i, y_i\}_{i=1}^{N}$ be a training dataset of N samples, and $\hat{\ell}(f^m)$ 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 upper bounded by:

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

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

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

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

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

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

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

Since both strategies use the same model architecture, the empirical loss remains unchanged:

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

Additionally, dynamic fusion tends to assign higher weights to models with lower complexity. Therefore,

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

Since the confidence term $M\sqrt{\frac{\ln(1/\delta)}{2N}}$ is independent of the fusion 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 fusion yields a tighter generalization error bound than static fusion.

Remark. To further explore the nature of dynamic collaboration, we analyze the generalization error upper bound in multimodal systems as shown in Eq 7:

$$\begin{aligned} GE(f) \leq\ & |\mathcal{M}| \left(\mathcal{R}_N(\mathcal{H}) + \sqrt{\frac{\ln(1/\Delta)}{2N}} \right) + \sum_{m=1}^{|\mathcal{M}|} e\hat{r}r(f^m) \\ & - \sum_{m=1}^{|\mathcal{M}|} \left[\frac{1}{|\mathcal{M}|} \underbrace{Cov(\omega^m, \ell^m)}_{\textit{Negative self-correlation in DSC}} - \frac{|\mathcal{M}| - 1}{|\mathcal{M}|} \sum_{j \neq m} \underbrace{Cov(\omega^m, \ell^j)}_{\textit{Positive collaborative correlation in DSC}} \right]. \end{aligned} \tag{7}$$

This bound reveals that the participation weight ω^m of a model is not only negatively correlated with its own loss (self-regularization), but also positively influenced by the losses of its collaborators. The first covariance term penalizes poor performance (negative self-correlation), while the second encourages collaboration with underperforming agents (positive cross-covariance). Such behavior closely mirrors human-like cultural learning, where individuals adapt based on both personal and peer performance, thus validating the dynamic collaboration mechanism in DISC.