(a) | (b) Discovery Learning Stage **Prompt Generation Stage** Dataset file: CLIP. " a photo of Text a cattle, Regularization for Encoder $\hat{\pmb{z}}_i^t$ which is a predictions: type of large omnivores and herbivores. " $H(\cdot)$, Cattle \mathcal{L}_{Dist} $Sinkhorn(\cdot)$ Prompt ‡sg. Generation Net. z'_{i}^{t} view_1 \mathcal{L}_{CMC} Self-supervised Multimodal **Multimodal** learning for **Fusion Fusion** features: Prototypes z'^{v}_{i} z_i^v $\mathcal{L}_{res} = \mathcal{L}_{res}^{u} + \mathcal{L}_{res}^{s}$ view_2 z'^t_i CLIP. Text Meta Module Encoder $\boldsymbol{z^t}$ Share weight CLIP. Vision Encoder 紫 z^v (view_1,2) view 1 view 2