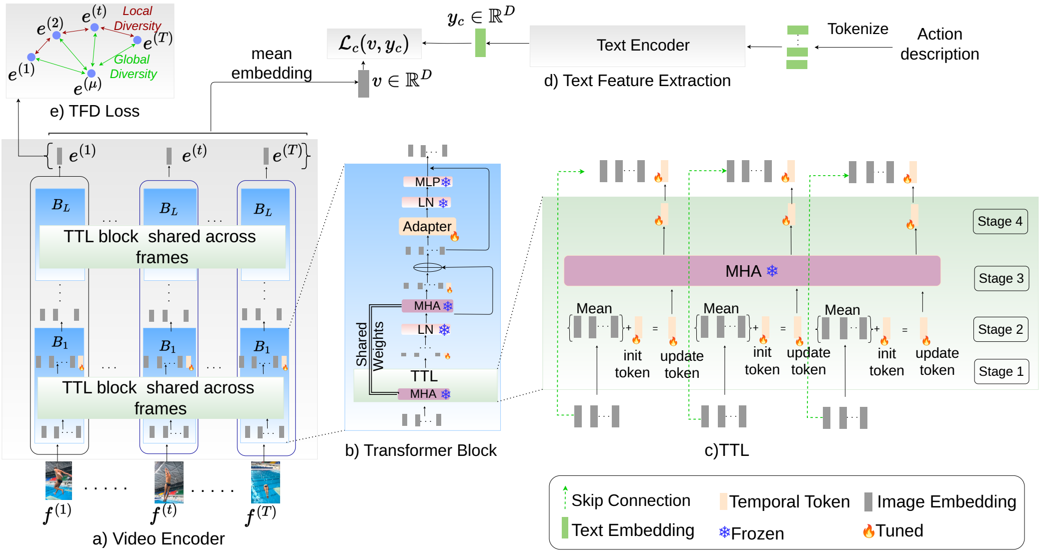

Figure 1: ***Overview of the proposed method:*** T2L utilizes Temporal Token Learning (TTL) and Temporal Feature Diversity Loss (TFD) to efficiently learn temporal aspects, eliminating the need for the frame integration module, a bottleneck in adapting image models for video understanding. Module (a) depicts the architecture of a video encoder, where each frame is parallel encoded by a vision transformer. Within the vision transformer, each block is adapted using spatial adapters shown in (b). Temporal relations between frames are learned by module (c) - Temporal Token Learning. Module (d) shows the text encoder, and module (e) shows the Temporal Feature Diversity Loss $\mathcal{L}_{TFD}$).