**Explanation of Fig. 3:**

Figure 3 illustrates how the system captures high-level activity patterns by analyzing spatio-temporal relationships between groups of STIP features. Each node in the graph (e.g., VaV\_aVa​, VbV\_bVb​, VcV\_cVc​) represents a cluster of STIPs within a segment that are labeled with the same activity class, such as "Exit Car" or "Open Trunk." The edges between these nodes (e.g., Ψab\Psi\_{ab}Ψab​) do not reflect individual STIP-to-STIP connections but instead summarize how frequently and closely these activity classes co-occur across training segments. These edge weights are computed using a combination of spatial and temporal Gaussian models along with co-occurrence statistics. As a result, the graph encodes common patterns of event progression (e.g., exiting a car is often followed by opening the trunk), providing a compact and meaningful structure that represents the dynamics of the video.