Joint video-language modeling has been attracting increasing attention in recent years, signifying a return to early AI goals of cooperative cognitive systems. However, many approaches fail to leverage the complementarity across vision and language. For example, they may rely on a fixed visual model or fail to leverage the underlying compositional semantics inherent in language. In this talk, I will discuss a sequence of recent work in my group that indeed directly and holistically models vision and language in ways that not only jointly models the visual and the lingual signals but also exploits the compositionality in language to learn better representations. The first method I will discuss jointly embeds a deep video model and a compositional text model that sits on a dependency-tree structure. The joint embedding fine-tunes all three model components together under a unified cost function and affords three tasks: text generation, text retrieval and video retrieval. The second method I will discuss explicitly relates visual and speech signals in a bimodal sparse model. The bimodal model represents visual and speech signals in separate but linked dictionaries facilitating a bidirectional generative capability. Furthermore, we enforce a structure to the dictionaries that captures the compositionality of the underlying spoken language. Both approaches capture visual and lingual signals from the bottom-up and demonstrate the potential of signal-level cross-modal embeddings for realizing next generation cooperative cognitive systems.