Introduction

Definition

VLMs learn to map the relationships between text data and visual data such as images or videos, allowing these models to generate text from visual input or understand natural language commands in the context of visual information[1].

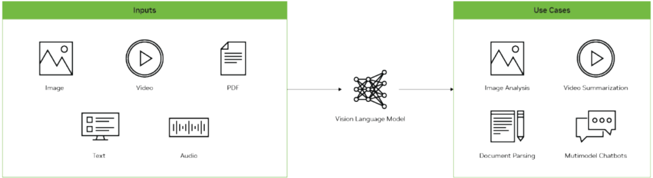

Figure 2: Visualization of a Vision-Language Model

[1] https://www.ibm.com/think/topics/vision-language-models