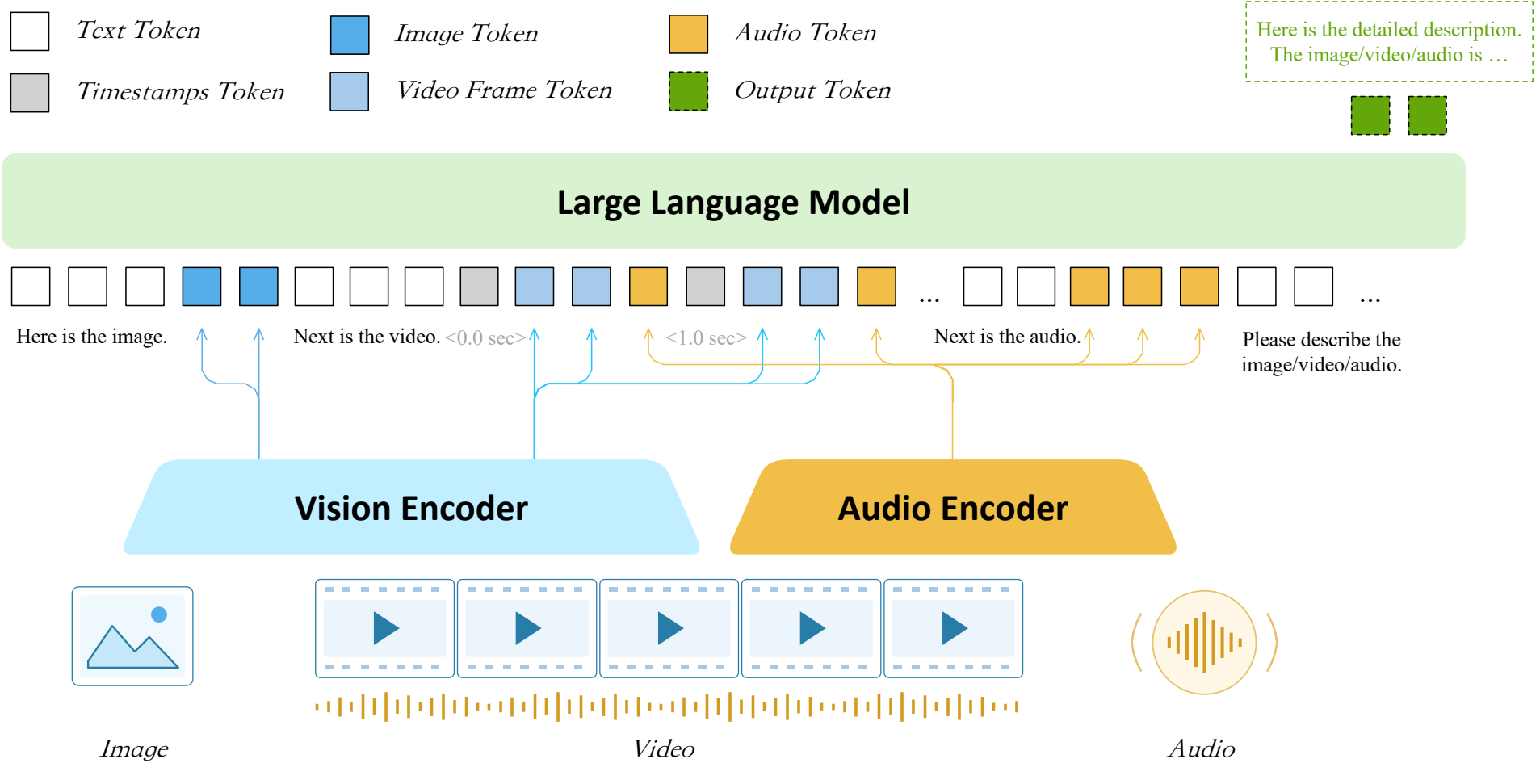

Text Token
Image Token
Audio Token
Timestamps Token
Video Frame Token
Output Token
Here is the detailed description. The image/video/audio is …
Large Language Model
Here is the image.
Next is the video. <0.0 sec>
<1.0 sec>
Next is the audio.
Please describe the image/video/audio.
Vision Encoder
Audio Encoder
Image
Video
Audio