

Notes on multimodal fusion

**Image captioning with Meta’s Llama**

* Combining computer vision with natural language processing creates powerful tools for understanding visual content.
* Three main stages of the image captioning process with a multimodal large language model (LLM) are:
  + Input processing
  + Image validation and encoding
  + Multimodal LLM processing
* Input processing receives and prepares the image and optional text prompt.
* Image validation and encoding validate and convert the image into a format (e.g., Base64) suitable for the model.
* Multimodal LLM processing combines visual and textual information to generate a descriptive caption.
* Core components of the image captioning system to produce captions tailored to prompts are:
  + Visual encoders
  + Text embedding
  + Fusion layers
  + Language generation tools
* Implementing an image captioning system using Meta’s Llama 4 Maverick model via IBM watsonx involves:
  + Importing libraries and authenticating access
  + Encoding images and preparing prompts
  + Sending combined image-text messages to the model
  + Extracting descriptive text from the model’s response

**Text-to-video generation with OpenAI’s Sora**

* Sora is a multimodal, diffusion-based transformer model developed by OpenAI that can generate high-quality video from text or image inputs.
* For accurate results, you must craft a structured prompt and include essential elements, such as scene context, visual details, and motion required in your clip.
* Steps for text-to-video generation:
  + Open your browser and go to sora.openai.com to access the official Sora interface.
  + If not logged in, click “Log In” or “Sign Up” for a new OpenAI account.
  + After logging in, you’ll land on the “Explore page,” where you can browse others’ videos for inspiration.
  + Use the composer at the bottom to enter your text prompt describing the video you want.
  + Before creating, review your settings:
    - Choose Type: Video
    - Set aspect ratio, resolution, duration, number of variations, and style preset
    - Options depend on your OpenAI subscription tier
  + Click “Create video” to submit your request; processing takes 30 seconds to a few minutes
    - Finished videos appear under “My Videos”
    - Hover to “Preview”
    - Click to open in a lightbox and use the arrow keys to view variations
* Select a variation to refine using the editing toolbar:
  + “Edit” storyboard or recut clips
  + Use “Remix” to describe changes in natural language
  + Use “Blend” to merge with another video
  + Use “Loop” to create seamless repeats
* After editing, a new variation is added to your set.