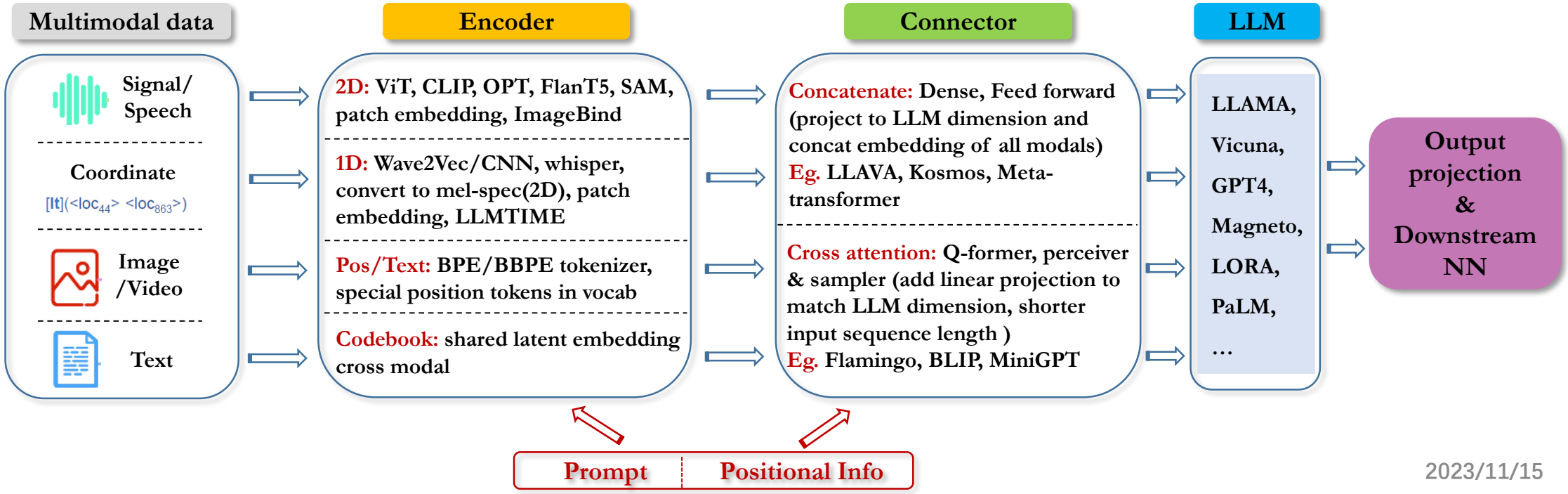
Notes:

- ✓ '**Connector**' may described as bridging/alignment/adapter/neck/general purpose interface in different paper, but with same goal/role:
  1. Fuse features of different modality
  2. Dimension matching with LLM (often with linear projection, which is proven to be effective empirically)
- ✓ Order of **Encoder** and **Connector** can be reversed, for example,
  1. In Meta-transformer, firstly connector/data-to-sequence tokenizer embedded different modalities, the output is then fed into encoder to get latent embedding
  2. In NExT-GPT, firstly different encoders are adopted to encode raw input, the output is then fed into connector.
- ✓ Patch embedding ←→ linear projection
- ✓ References: Flamingo, Kosmos-1/2, MiniGPT-4, BLIP-2, LLAVA, LaVIN, Meta-transformer, NExT-GPT, mPLUG-Owl-1/2, UniCode, ImageBind, AdaLink, …
- ✓ Training skills: contrastive learning, InfoNCE loss, CRPS loss, …



2023/11/15