Table 2: Advanced Multimodal Tasks

Task Type	**Description**	**Strong VLM Models**
Image Captioning	Generates a text description for the image	**BLIP**, **BLIP-2**, **GIT**, **MiniGPT-4**
Visual Question Answering (VQA)	Answers questions based on the image content	**BLIP-2**, **Flamingo**, **MiniGPT-4**, **LLaVA**
Visual Grounding	Locates regions in the image based on a textual description	**Grounding DINO**, **OWL-ViT**, **GLIP**
Text-to-Image Retrieval	Finds matching images from a text query (or vice versa)	**CLIP**, **ALIGN**, **Florence**
Multimodal Reasoning	Performs reasoning using both image and language modalities	**GPT-4V**, **Kosmos-2**, **MiniGPT-4**, **LLaVA**