Model / Variant	Multimodal?	What Modalities Supported	Model tested on 16gb ram
Base Qwen / Qwen2 / Qwen2.5 (LL-only)	**No**	Text only	These are standard LLMs; do not accept images or audio input.
Qwen-VL / Qwen2-VL	**Yes**	Vision + Text (images + text inputs)	Tested on ollama. Model is qwen 2.5vl:7b and output is very slow.
Qwen-Audio / Qwen2-Audio	**Yes**	Audio (speech or sound) + Text	Tested on ollama. Model is 7b and output is very slow.
Qwen2.5-Omni	**Yes**	Text, Images, Audio, Video inputs; outputs: Text & Speech (audio) in streaming manner.	Tested on huggingface locally, output is comparatively faster.
Qwen3	Partial multimodal (depends on variant)	The VL versions / variants would support vision + text; but many of the Qwen3 models are LL-only.	The models available are mostly text based and new multimodal models are not available yet.