S rostoucí produkcí multimédií se zvyšuje také poptávka po efektivních metodách jejich vyhledávání. Jedním z klíčových úkolů v této oblasti je vyhledávání konkrétních položek ve velkých nestrukturovaných kolekcích obrázků pomocí textových dotazů. V posledních letech tomuto oboru dominují hluboké neuronové sítě, které jsou trénovány na mapování obrázků a textu do joint-embedding prostoru. V rámci našeho výzkumu jsme porovnali výkon několika předtrénovaných sítí, včetně CLIP, OpenCLIP, ALIGN a BLIP2, na různých datasetech. Kromě toho jsme zkoumali, jak množství informací obsažených v textových dotazech ovlivňuje výkon modelů. Hodnotili jsme také konzistenci vnímání podobnosti mezi obrázky modely a lidskými posudky. Naše zjištění ukazují, že modely OpenCLIP se osvědčují jako velmi účinné při vyhledávání konkrétních položek pomocí dotazů a dobře se shodují s lidským vnímáním podobnosti. Dále jsme zjistili, že podrobnější informace v textových dotazech zlepšují výkon modelů.