As the production of multimedia continues to grow, the demand for effective multimedia retrieval methods increases. One critical task in this domain is known-item search within large unstructured collections of images using text queries. In recent years, this field has been dominated by deep networks trained to map both images and text in joint-embedding space. We evaluated multiple pre-trained networks, including CLIP, OpenCLIPs, ALIGN, and BLIP2, comparing their performance across various datasets. Additionally, we investigated how the amount of information provided within the text queries influences model performance. We also assessed the consistency of models' perceived image-image similarity with human judgments. Our findings indicate that OpenCLIP models excel in known-item search with queries and align well with human perception of similarity. Furthermore, we observed that providing more detailed information in text queries enhances model performance.