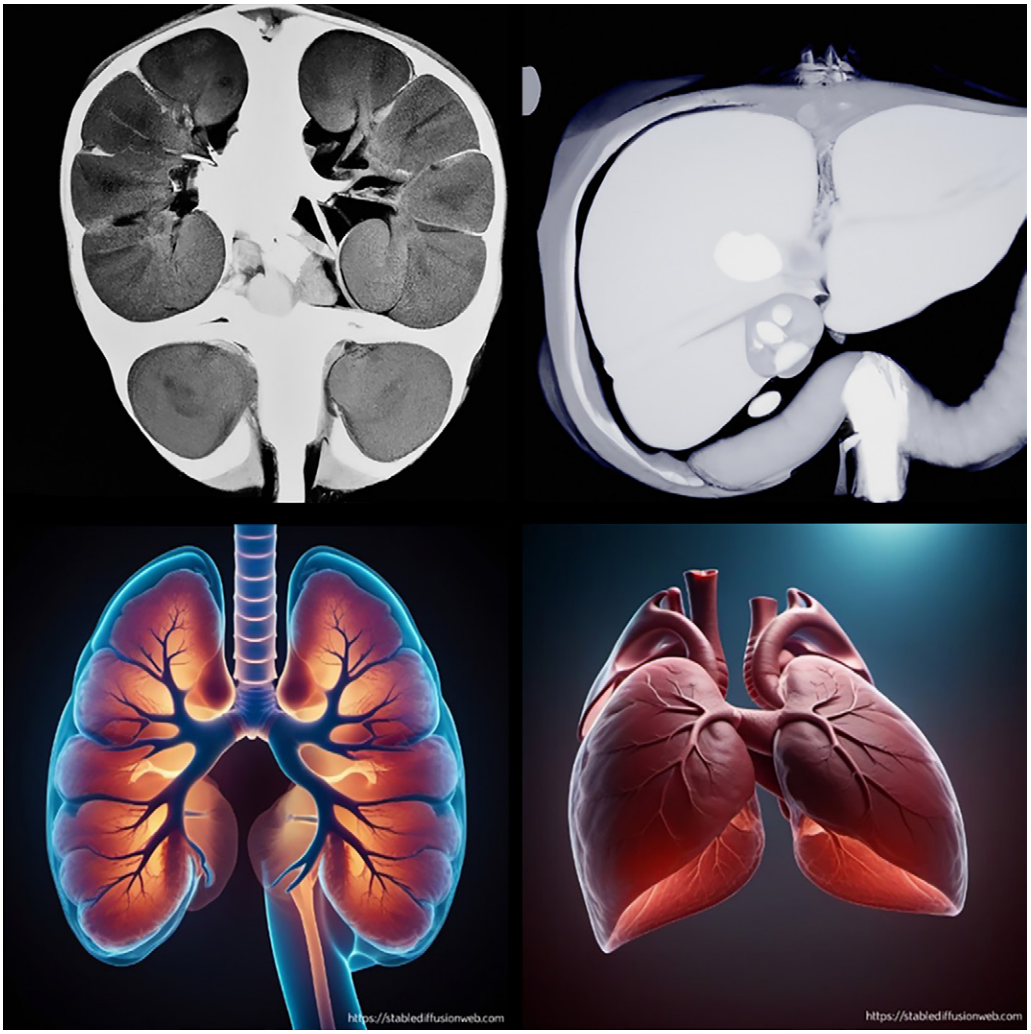
Deep Learning Models Connecting Images and Text: A Primer for Radiologists

An Ni Wu, MD • Merve Kulbay, MD, PhD* • Phillip M. Cheng, MD, MS • Alexandre Cadrin-Chênevert, MD, BIng • Laurent Létourneau-Guillon, MD, MSc Gabriel Chartrand, PhD • Jaron Chong, MD • Emmanuel Montagnon, PhD • Ismail Ben Ayed, PhD • An Tang, MD, MSc*

* A.N.W. and M.K. contributed equally to this work.
Author affiliations, funding, and conflicts of interest are listed at the end of this article.


https://stablediffusionweb.com
https://stablediffusionweb.com

In radiology practice, medical images are described and interpreted by radiologists in text reports. Recent technical developments enabling deep learning models to connect images and text may facilitate the radiologic workflow. These developments include advances in data embedding, self-supervised learning, zero-shot learning, and transformer-based model architectures. Models connecting images and text can be divided into four categories: *(a)* Text-image alignment models associate text descriptions with corresponding images. *(b)* Image-to-text models create text descriptions from images. *(c)* Text-to-image models generate images from text descriptions. *(d)* Multimodal models integrate and interpret multiple types of data such as images, videos, text, and numbers simultaneously. Potential clinical applications of these models include automated captioning of medical images, generation of the preliminary radiology report, and creation of educational images. These advances may enable case prioritization, streamlining of clinical workflows, and improvements in diagnostic accuracy.

INFORMATICS

RadioGraphics 2025; 45(9):e240103
https://doi.org/10.1148/rg.240103

Content Code: IN

Abbreviations: AI = artificial intelligence, BLIP = Bootstrapping Language-Image Pre-training, CLIP = Contrastive Language-Image Pre-training, LLaVA = Large Language-and-Vision Assistant, ViT = Vision Transformer

TEACHING POINTS

- Deep learning models connecting images and text are categorizable by their inputs and outputs. Text-image alignment models associate text with corresponding images. Image-to-text models create text from images, while text-to-image models generate images from text. Multimodal models integrate and interpret multiple types of data simultaneously
- Deep learning models can be broadly categorized based on the nature of the tasks they solve into two types: discriminative or generative models.
- Zero-shot learning offers a promising alternative to task-specific learning by enabling models to classify new, unseen classes without requiring explicit examples during training.
- Transformer models have been proven to be both effective and efficient in understanding the meaning of words in their larger contexts and are now pervasive in natural language processing applications.
- Most models linking images and text are trained on internet data, making them successful in general domains. However, models must be fine-tuned with medical data to be suitable for biomedical applications.

Introduction

Analysis and interpretation of medical images rely on the expertise of radiologists. Medical imaging studies, along with their corresponding radiology reports, contain a wealth of information essential for clinical management. The increasing number of imaging studies presents a substantial challenge for radiologists, who must manage a growing volume of interpretations while meeting the demand for rapid turnaround times. This situation places considerable pressure on radiologists to maintain high productivity and accuracy (1).

To address these challenges, technology may assist radiologists by enhancing their efficiency and accuracy. For example, computer-aided detection systems help radiologists detect and diagnose abnormalities (2–4). However, conventional computer-aided systems are often limited to specific tasks and their performance can degrade when applied to different datasets or tasks, requiring datasets curated by experts for effective training (5).

Deep learning models, a type of artificial intelligence (AI), use neural networks with multiple layers to learn complex patterns from large datasets. Recent technical advances in computer vision and natural language processing, along with the wider availability of data and increases in computing power, have enabled models to connect images and text (6–11). In recent studies (11–13), medical image captioning using models pretrained on large and diverse datasets has been found to have high accuracy. These models could adapt to various new applications such as generating preliminary radiology reports from medical images (Fig 1) (14). In addition, multimodal models may integrate multiple types of data to inform patient care (15,16). The use of models connecting images and text may transform how radiologists review imaging examinations and report diagnostic findings (17–20).

This article includes a discussion of recent technical developments in deep learning that have enabled models connecting images and text. A glossary of commonly used terms in deep learning is provided (Table). Also discussed are differences between discriminative and generative models. Next, essential concepts of data embedding, self-supervised learning, and zero-shot learning are summarized. Then, four categories of models connecting images and text are discussed: *(a)* text-image alignment models, *(b)* image-to-text models, *(c)* text-to-image models, and *(d)* multimodal models. For each type of model, discussions of its general architecture, summaries of the underlying key concepts, and descriptions of potential applications are included.

Overview of Deep Learning Models Connecting Images and Text

Deep learning models connecting images and text are categorizable by their inputs and outputs. Text-image alignment models associate text with corresponding images. Image-to-text models create text from images, while text-to-image models generate images from text. Multimodal models integrate and interpret multiple types of data simultaneously (Fig 2). In this article, the authors focus on models that process image and text data, although multimodal models may integrate a wider variety of data types including video, audio, frequency, and sequencing data.

Models taking images as input may be most relevant for radiologists in clinical settings (eg, for preliminary generation of a radiology report) (14), whereas models taking text as input may be useful for educational and research purposes (eg, prompts generating synthetic examples) (Fig 3) (21).

Discriminative and Generative Models

Deep learning models can be broadly categorized based on the nature of the tasks they solve into two types: discriminative or generative models (Fig 4). Discriminative models learn the decision boundaries between different classes in the input data. Their primary goal is to determine which class new input data belong to based on learned features. For example, when presented with a new image, a discriminative model may decide whether it contains a liver or a kidney. In contrast, generative models learn the underlying patterns of the training data, from which they generate new examples (22). Thus, generative models provide a mechanism for creating new text or image data.

Data Embeddings

Embedding Space

Deep learning models connecting images and text are trained on images paired with relevant text descriptions. Training data are represented as lists of numbers (ie, vectors) called data embeddings. The numerical values of an embedding encode relevant concepts in the data, effectively locating data items in an “embedding space.” The embedding space can be considered as a map, where the relative positions of different locations can convey semantic relationships.

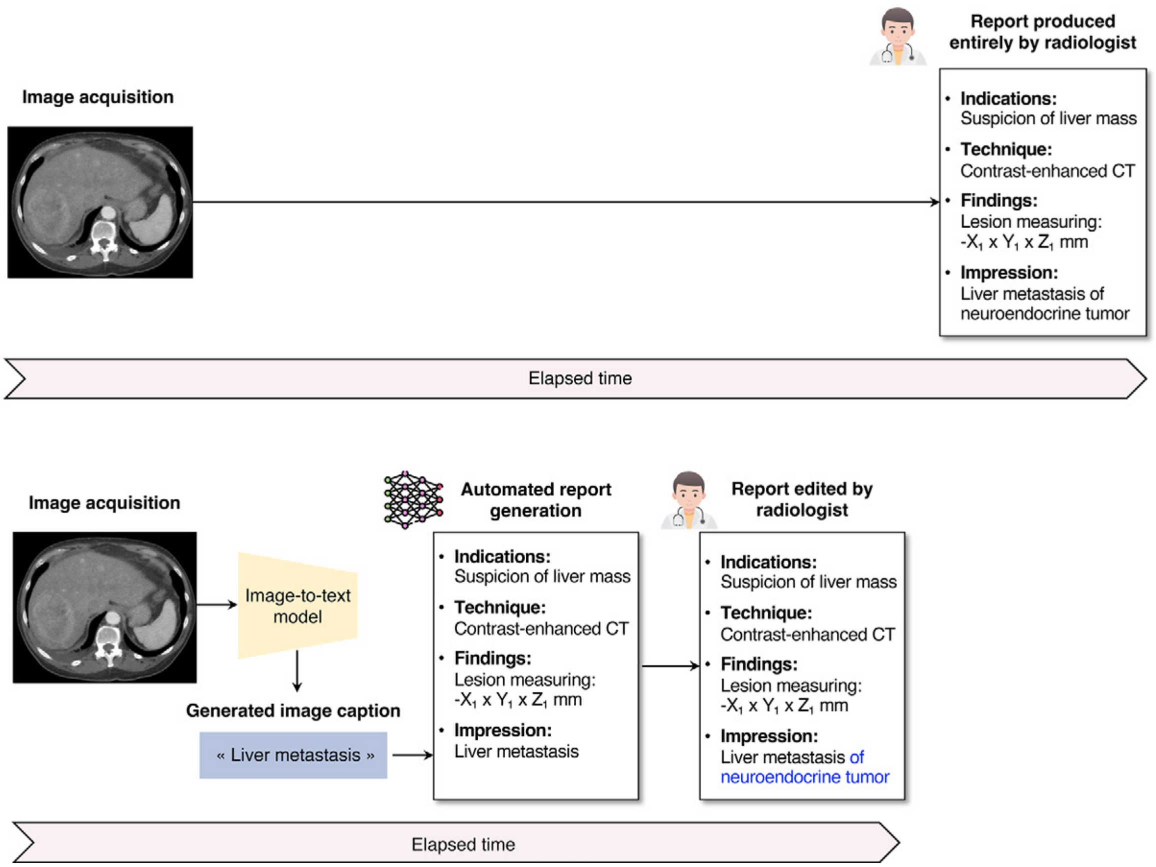

Report produced entirely by radiologist
Image acquisition
• Indications: Suspicion of liver mass
• Technique: Contrast-enhanced CT
• Findings: Lesion measuring: -X_1 x Y_1 x Z_1 mm
• Impression: Liver metastasis of neuroendocrine tumor
Elapsed time
Image acquisition
Automated report generation
Report edited by radiologist
Image-to-text model
Generated image caption
« Liver metastasis »
• Indications: Suspicion of liver mass
• Technique: Contrast-enhanced CT
• Findings: Lesion measuring: -X_1 x Y_1 x Z_1 mm
• Impression: Liver metastasis
• Indications: Suspicion of liver mass
• Technique: Contrast-enhanced CT
• Findings: Lesion measuring: -X_1 x Y_1 x Z_1 mm
• Impression: Liver metastasis of neuroendocrine tumor
Elapsed time

Figure 1. Illustration shows the concept of automated report generation with image-to-text models. The current interpretation workflow relies on the production of reports by radiologists. Image-to-text models may assist radiologists by analyzing the medical images and automatically populating sections of radiology reports in the background. The radiologist then interprets the images and edits the generated reports to produce a final report. If an AI-augmented workflow becomes efficient, it has the potential to save interpretation time compared with that of the existing workflow. (Icons made by Freepik, Amethyst Design from www.flaticon.com.)

Practically speaking, during the embedding process, a deep learning model extracts the most relevant information from an input image or text and compresses it into a series of numbers. The meanings and concepts extracted by the model are represented in how big or small these numbers are, which reflects their relationships. These collections of numbers are known as embedding vectors. Embeddings can be generated with specialized deep learning models, or they can be learned by a subset of a deep learning model during training.

Text Embeddings

In a word embedding space, the locations of word vectors are determined by their meanings. In effect, similar words are located close to each other in this embedding space (Fig 5A) (23). To form word embeddings, the source text is processed into an analyzable form before being converted by deep learning models into embedding vectors (Fig 5B) (24). These embedding models have been trained on extensive text data to learn useful insights and underlying patterns.

Image Embeddings

Similarly, image data can be converted to embeddings that summarize aspects of the image content. For example, medical images that depict the same organ are organized within the same cluster, while dissimilar images are more distant from each other (Fig 6A). Image embeddings are also generated by deep learning models (Fig 6B) (23). In particular, convolutional neural networks are especially useful for extracting features such as edges, textures, and shapes that allow recognition of an organ or lesion (25,26).

Shared Embedding Space

In models connecting images and text, image and text embeddings are aligned so that a model can associate visual concepts with their textual representations. Since embedding models are typically trained from large nonspecialized images and datasets, incorrect conceptual relationships might arise when these models are used to generate embeddings for domain-specific applications (27). For example, a dataset of radiology text reports and corresponding images may contain similar concepts such as “fatty liver” and “steatosis.” A model without specific domain knowledge may fail to interpret these phrases as referring to similar medical image findings. This highlights the importance of domain-specific knowledge for specialized concepts.

To address this challenge, semantic categories curated by domain experts (ie, specific labels annotated by domain experts that tell the category to which each example belongs)

Glossary of Commonly Used Terms in Deep Learning

Term	Definition
Computer vision	A field of AI enabling computers to understand images and videos
Contrastive learning	A learning approach in which a model represents similar data close together and dissimilar data far apart
Decoder	A part of a neural network that reconstructs the output of an encoder into the desired output. Decoders may convert vectors of numbers into images or text.
Deep learning	A type of learning in which the algorithm learns a hierarchy of features in the data using neural networks with many layers
Discriminative models	A type of model that focuses on learning the boundaries separating different classes
Embedding	A compressed numerical representation of the relevant information from an input image or text
Embedding space	A map in which embeddings are represented so that their relative positions capture their semantic relationships
Encoder	A part of a neural network that converts input data into a numerical representation. Encoders may convert images or text into vectors of numbers.
Encoder-decoder models	A type of model that includes an encoder and a decoder. This type of model may transform one type of data (image or text) into another (image or text).
Generative models	A type of model that learns patterns to generate new samples similar to the original training data
Image-to-text models	A type of generative model designed to create text descriptions from images
Model	An algorithm that learns from training data how to perform specific tasks on new data
Multimodal models	A type of deep learning model that integrates multiple types of data (eg, images, videos, text, and numbers) into a unified framework, allowing it to interpret all these types of data simultaneously
Natural language processing	A branch of AI enabling computers to comprehend, generate, and manipulate human language
Neural networks	A model inspired by the structure of biologic neurons composed of multiple layers of connected nodes. Modern neural networks may contain thousands to millions of nodes.
Noisy labels	Incorrect or misleading annotations that negatively affect the performance of a deep learning model
Self-attention	A mechanism that allows the model to focus on different parts of the input data by weighing their importance (42)
Self-supervised learning	A type of learning from unlabeled data relying on careful design of tasks to produce labels from existing information within the data; for example, a model may be trained to predict masked parts of an image
Semantic relationship	Connection between concepts (whether words or images) based on their meaning; in an embedding space, concepts with similar meanings are located close together
Supervised learning	A type of learning where all training data are explicitly labeled
Text-image alignment models	A type of model that associates text descriptions with corresponding images by learning a shared embedding space
Text-to-image models	A type of generative model designed to create images from text descriptions
Transformer	An efficient and scalable neural network architecture widely used in natural language processing, using self-attention to understand the meaning of words in their larger context (42). Transformers have more recently been applied to images (see vision transformer).
Vision Transformer	An adaptation of the transformer architecture to images, where an input image is divided into a sequence of patches treated in a similar way to words in a natural language processing application (44)
Zero-shot learning	A technique where a model can identify a concept that was not explicitly observed during training; for example, zero-shot learning trained on "calcified lung granulomas" and "normal spleens" may recognize that splenic calcifications may represent "calcified splenic granulomas."

can provide additional context to support the appropriate organization of the embedding space (27). In this example, an expert could indicate that image descriptions of a fatty liver and steatosis in reports refer to identical concepts. This way, the trained model can learn to recognize these phrases as referring to similar findings.

Self-supervised Learning

Self-supervised learning refers to a process by which a model can learn concepts from unlabeled data through carefully designed tasks that exploit existing information within the data (28). These tasks are called pretext tasks and do not require explicit labeling. An example of a pretext task is contrastive learning, in which a model learns to represent data in such a way that similar data are close together while dissimilar data are far apart (28).

For instance, consider a liver image and a kidney image; several slightly altered versions of these images can be generated using a series of transformations. A model can then be trained to recognize the modified versions of the liver image as similar to each other and to distinguish these liver images from the modified versions of the kidney image. In other words, the goal of training is to maximize the similarity between similar image pairs (ie, the liver images should be represented similarly, and the kidney images should be

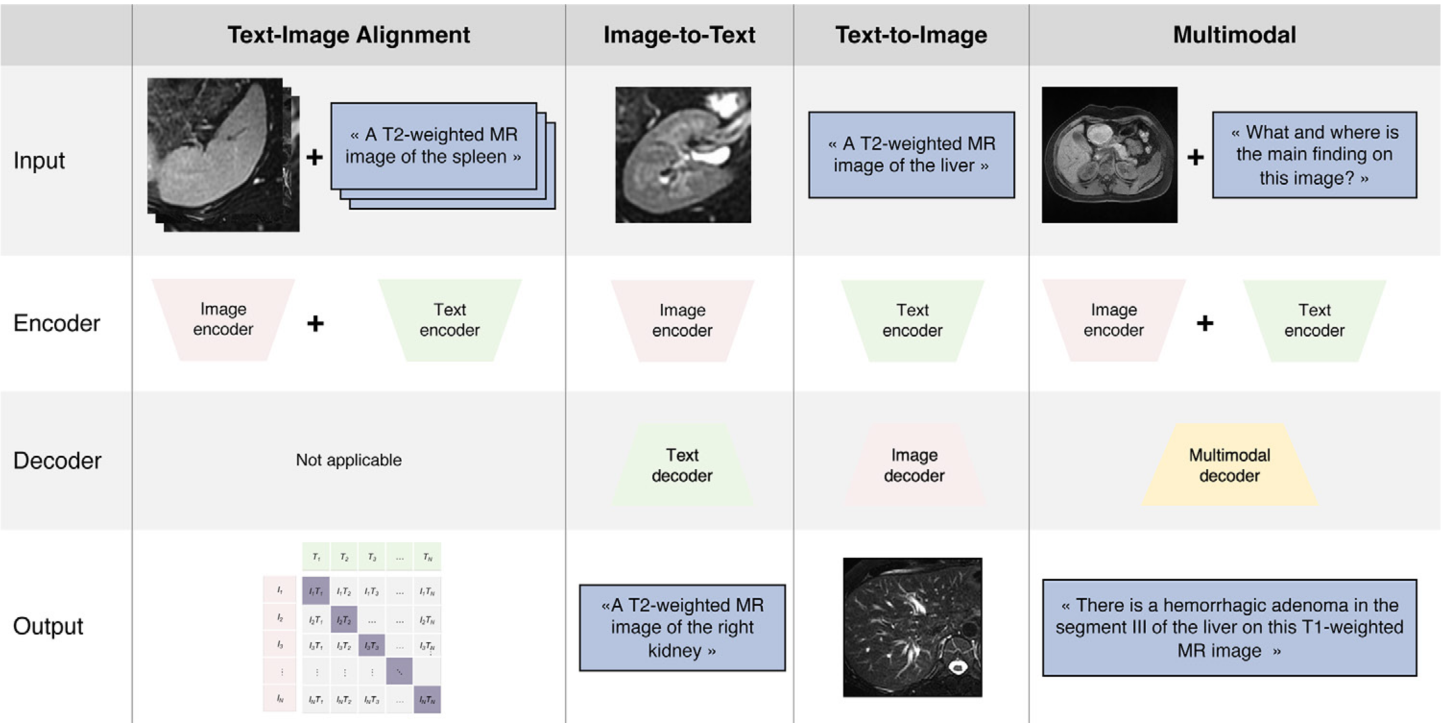

Text-Image Alignment
Image-to-Text
Text-to-Image
Multimodal
Input
« A T2-weighted MR image of the spleen »
« A T2-weighted MR image of the liver »
« What and where is the main finding on this image? »
Encoder
Image encoder
Text encoder
Decoder
Not applicable
Text decoder
Image decoder
Multimodal decoder
Output
«A T2-weighted MR image of the right kidney »
« There is a hemorrhagic adenoma in the segment III of the liver on this T1-weighted MR image »

Figure 2. Summary of the four categories of deep learning models connecting text and images. These models are text-image alignment, text-to-image, image-to-text, and multimodal models. The input, encoder, decoder, and output for each model are illustrated. For text-image alignment, images and text are input into their respective encoders to produce an embedding for both images and text, with scores indicating their similarity (eg, how well they are aligned). In an image-to-text model, an image is fed into an image encoder, which produces features that are subsequently fed into a text decoder to generate text descriptions as output. In a text-to-image model, text descriptions are fed into a text encoder, which generates features that are subsequently fed into an image decoder to produce images as output. In a multimodal model, both images and text are fed into their respective encoders, are combined and processed through a joint embedding by a multimodal decoder, and produce integrated outputs such as text.

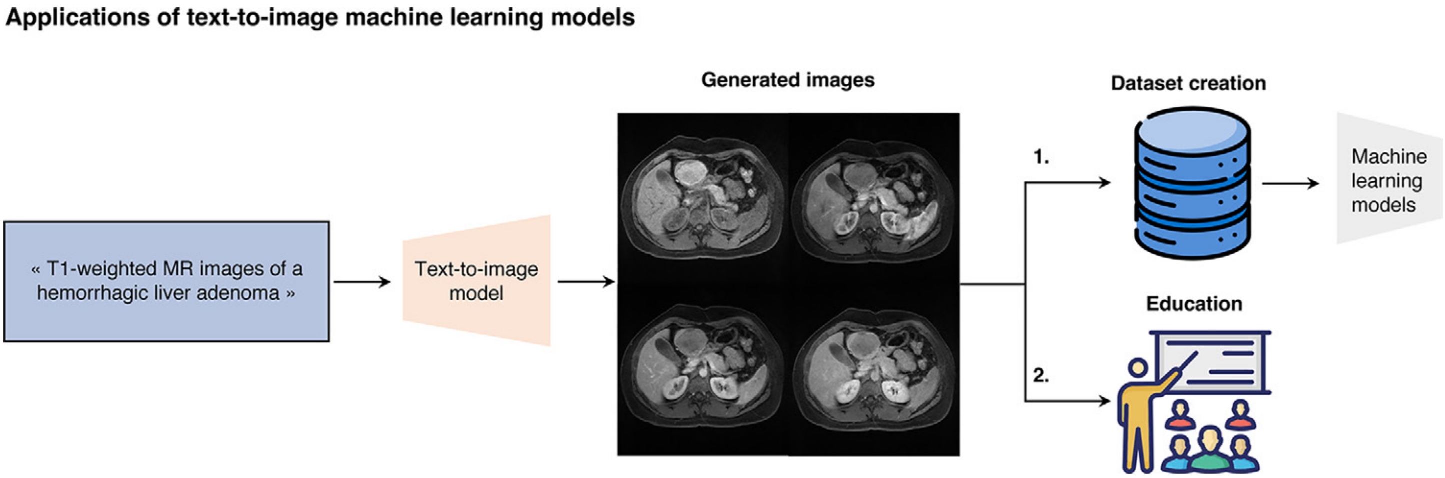

Applications of text-to-image machine learning models
« T1-weighted MR images of a hemorrhagic liver adenoma »
Text-to-image model
Generated images
1.
Dataset creation
Machine learning models
2.
Education

Figure 3. Illustration shows the leveraging challenges associated with the lack of domain-specific datasets and educational barriers in radiology with text-to-image models. Text-to-image models can augment existing radiologic datasets by generating image variations based on text descriptions. This approach could be useful when the datasets are small, as it provides additional data for training of AI models, reducing the time and cost to create large medical imaging datasets. These models could also be used to enhance training and education in radiology. Trainees could learn to associate text findings with corresponding radiologic images. (Icons made by Freepik, VectorPortal from www.flaticon.com.)

represented similarly) while minimizing the similarity between contrasting pairs (ie, the liver images should have representations distinct from those of the kidney images).

Other training approaches can be used to address the challenge of labeled data scarcity, such as transfer learning (29). Transfer learning involves using a pretrained model (ie, trained on a large but nonspecialized dataset) and adapting it to a new but related task with limited specialized data (30). The process of training a model on a new task is called fine- tuning. The idea of transfer learning is to leverage the knowledge gained by a model during pretraining to bootstrap its performance on another task with fewer data. Transfer learning is

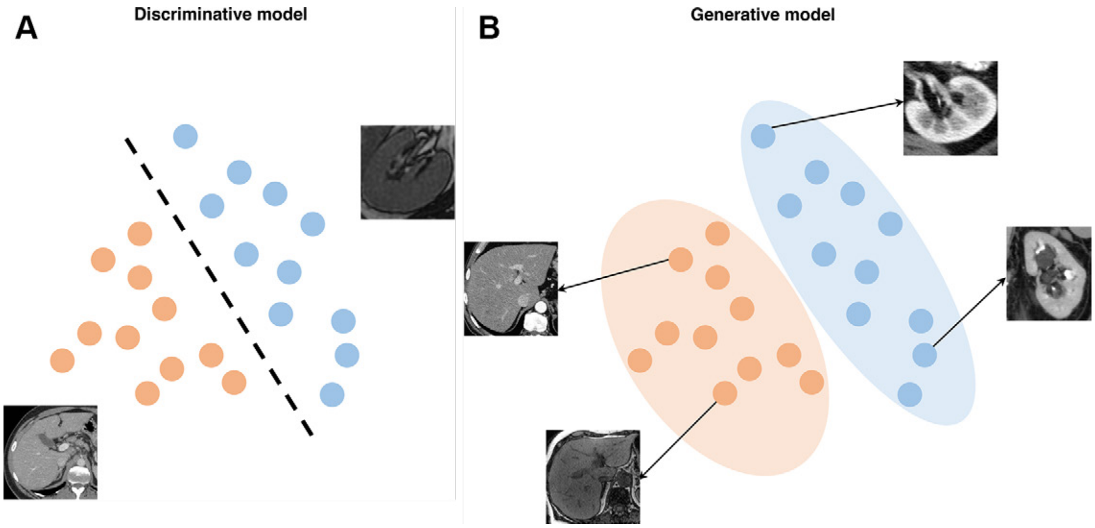


Figure 4. Illustrations show the differences between discriminative and generative models. **(A)** Discriminative models learn the decision boundary (represented by the dotted line) separating different classes in the training data. In this example, a discriminative model aims to find the optimal line that best separates the two classes (liver and kidney). The goal is to accurately classify new data points (ie, discriminative classification) by determining on which side of the boundary they fall, based on learned features (eg, to determine whether a new image contains a liver or a kidney). **(B)** Generative models learn the underlying patterns of the training data and generate similar new data. In this example, darker-colored small circles represent individual training examples of the liver (in beige) and the kidneys (in blue). The lighter-colored ellipses represent their respective distribution. The images in boxes illustrate the concept of synthetic (ie, generated) images.

useful in scenarios where robust pretrained models are accessible and when the new task is related to the original task.

Classifying Unseen Classes: Zero-Shot Learning

Task-specific learning requires a predefined and fixed set of classes. The number of classes is determined by the labels in the training data, and the objective of the model is to predict the correct class for new unseen inputs. For example, for the task of classifying images of organs into categories such as liver, kidney, or spleen, the model needs training examples for each of these classes (Fig 7A). Such models can achieve high accuracy if classes are well represented in the training dataset.

Most deep learning models developed for radiology have relied on this task-specific approach, which requires access to expert-labeled data. This reliance on labeled data means that generalization across the spectrum of abnormalities and image-acquisition techniques remains a substantial challenge for wider adoption (31). The scarcity of radiologists available to annotate images for purposes beyond patient care, along with the high costs and resources required for labeling radiologic images, limits the size of labeled datasets. For instance, even in the recent medical image segmentation decathlon (32), the size of datasets was typically limited to a few dozen or, at best, a few hundred CT or MR images.

Zero-shot learning offers a promising alternative to task-specific learning by enabling models to classify new, unseen classes without requiring explicit examples during training (Fig 7B) (33). This method reduces the need for extensive labeled datasets and improves the ability of the model to generalize across different tasks.

Models with zero-shot capabilities offer advantages in real-world applications, where the potential number of classes can change over time. Zero-shot learning allows models to classify new examples based on semantic attributes, which represent the essential characteristics of different classes and enable the model to recognize and classify concepts not encountered during training (Fig 8) (34,35). Classes are positioned as vectors in this space of attributes, with their location determined by how well they express these attributes (Fig 9) (33,36). This enables the model to make informed predictions about new classes based on their semantic similarities to known classes. Zero-shot learning is useful for models connecting images and text because it allows them to handle a broad range of concepts.

For instance, in a patient with an *Echinococcus* infection, hydatid cysts are most commonly found in the liver, while splenic hydatid cysts are rare. A radiologist may have never seen this condition in the spleen but could still recognize this diagnosis, drawing on their understanding of similar characteristics of this lesion in the liver and related medical knowledge. Zero-shot learning works similarly. Consider an AI model that has been extensively trained on medical images of common abdominal conditions such as hepatic hydatid cysts and simple cysts, but not on the rarer condition of splenic hydatid cysts. Despite this, with zero-shot learning, the model can still identify the splenic lesion as a potential hydatid cyst by leveraging its knowledge of spleen structures and pathologic features learned from other conditions.

Although zero-shot learning offers advantages in scenarios where acquiring labeled data is challenging, this technique still assumes that new classes of interest are adequately represented, even if they are not explicitly labeled in the training data. These new classes of interest effectively must belong to the same domain as that of the training data (31). A

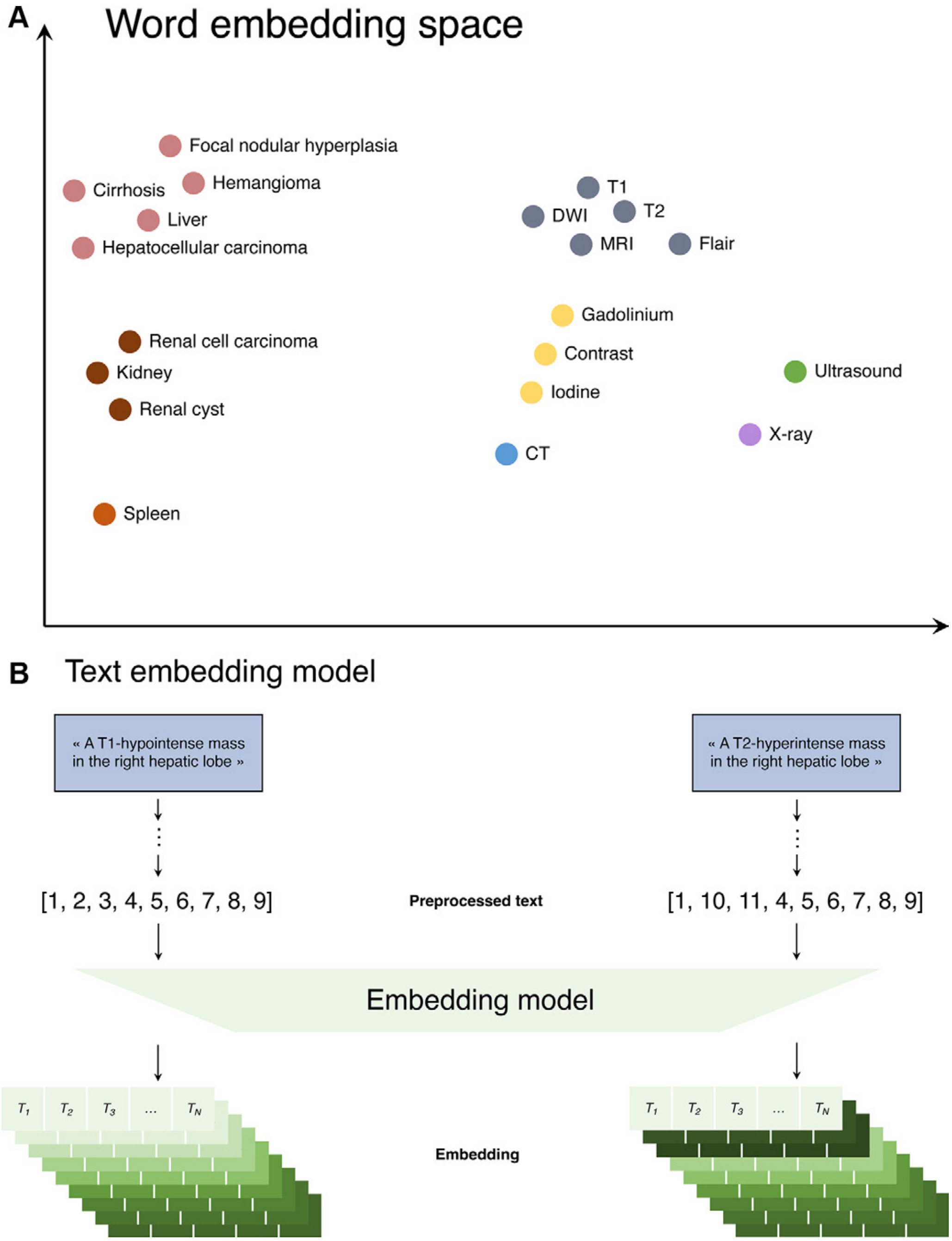


Figure 5. Illustrations show the word embedding space. **(A)** In a word embedding space, similar words are closer to each other, capturing their semantic similarity, while words with different meanings are positioned farther apart. The number of dimensions in the embedding space depends on the embedding model used. In this example, the space is limited to two dimensions, for simplicity. The X and Y axes in this space are new variables created by the embedding process, and they do not correspond to any specific, interpretable features of the words. **(B)** In word embedding, instead of using numbers that do not reflect relationships between words, the embedding model represents words in a more compact form by capturing their most relevant semantic relationships (23,24).

substantial domain shift between training and testing would make it difficult or impossible for the model to generalize. For example, if a model is trained only on a dataset of abdominal CT examinations and their radiology reports, it would not generalize to classify features in CT examinations of the head.

Model Architectures

Text-Image Alignment

Description.—Text-image alignment models link visual and textual information by associating the representations of images with the representations of their corresponding text descriptions. These models are crucial in generative tasks connecting text and images.

Developed by OpenAI in 2021, the Contrastive Language-Image Pretraining (CLIP) model is a text-image alignment model that has shown great generalizability in computer vision tasks and has gained broad interest in the medical field (17–20,37,38). The success of CLIP stems from its ability to learn in a self-supervised manner from massive amounts of text and image pairs publicly available on the internet, such as images with captions, without needing intensive human labeling efforts (39). CLIP has shown outstanding performance in zero-shot image classification predictions, where no labeled images are available for the target tasks (38).

Key Concepts.—CLIP, trained on a large dataset of images and their corresponding text descriptions, creates representations of text and images in a common embedding space. The

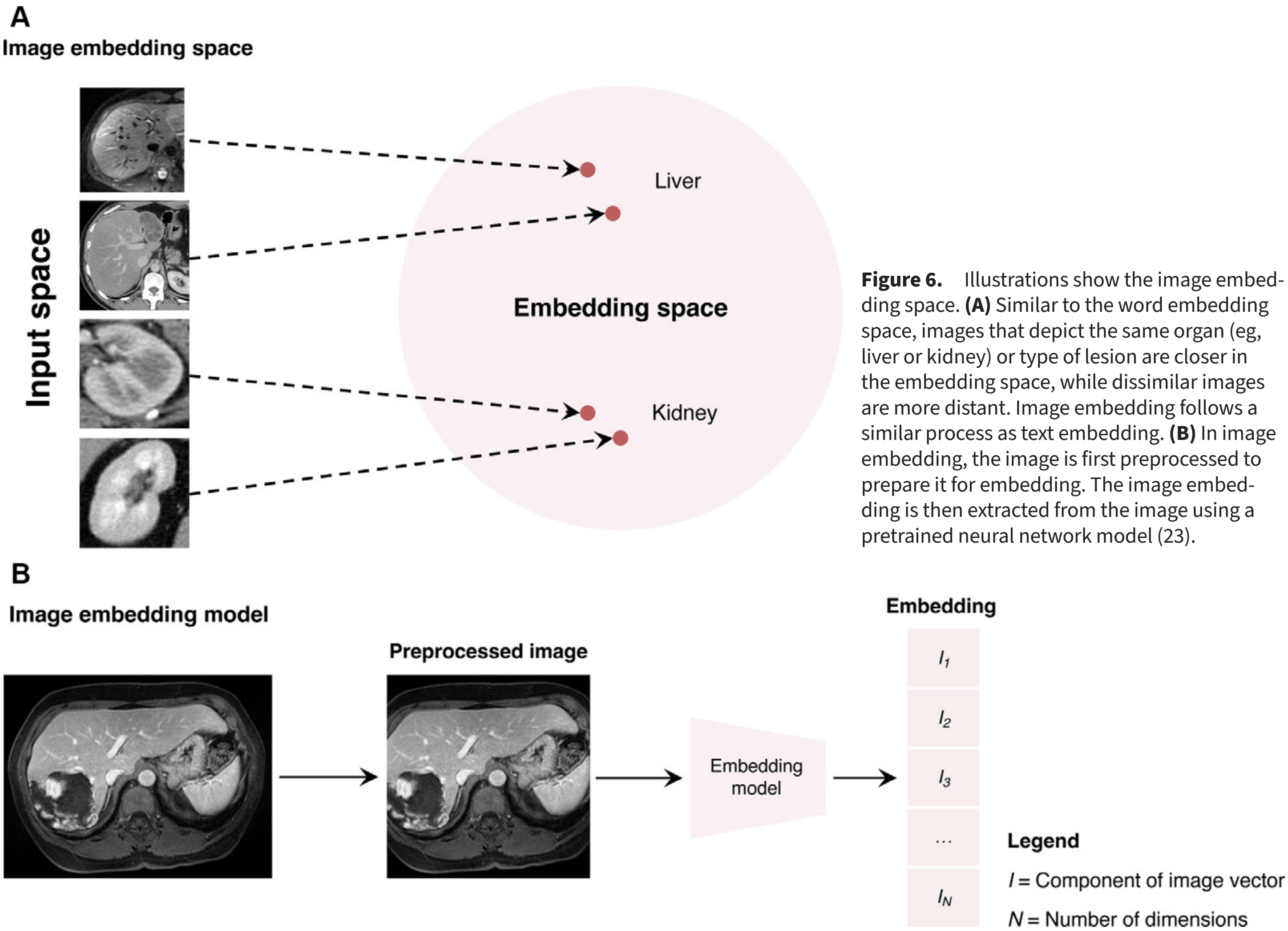


Figure 6. Illustrations show the image embedding space. **(A)** Similar to the word embedding space, images that depict the same organ (eg, liver or kidney) or type of lesion are closer in the embedding space, while dissimilar images are more distant. Image embedding follows a similar process as text embedding. **(B)** In image embedding, the image is first preprocessed to prepare it for embedding. The image embedding is then extracted from the image using a pretrained neural network model (23).

text descriptions may be phrased in natural language such as “A T2-weighted MR image of the liver.” The richness and diversity of the training labels are important to enable the zero-shot capabilities of CLIP (39).

CLIP could be viewed as a large-scale discriminative image classifier, with text captions serving as noisy labels. They provide implicit but imperfect guidance for associating images with text. Unlike typical labeled datasets, where each image is associated with a specific human-curated label, the captions in CLIP are naturally occurring text that accompanies images on the internet. They can vary in detail, accuracy, and relevance, often describing only parts of images or using ambiguous language. This variability introduces noise into the learning process, as the text does not always perfectly align with the visual content.

Despite this noise, CLIP is designed to learn robust associations between the image and text descriptions by leveraging a large-scale dataset of image-caption pairs. The noisy nature of the labels helps the model to generalize better as it learns to focus on the most relevant features across diverse and imperfect data. Through this self-supervised process, CLIP learns to create embeddings that capture the underlying relationships between images and their accompanying text, enabling it to perform tasks such as zero-shot image classification.

The architecture of CLIP combines a text encoder and an image encoder that create text and image embeddings in a shared embedding space. Training of the CLIP encoders is engineered to bring the embeddings of matching image-text pairs closer together in the embedding space, while pushing apart nonmatching pairs (Fig 10A). Once CLIP is trained, it can be used to sort images into classes based on text descriptions of the classes. The text encoder is used to convert these class descriptions into class embeddings. When an image needs to be classified, the image encoder generates its embedding, which is then compared with the class embeddings of interest using a similarity measure. The class with the highest similarity score is the predicted class for the image (Fig 10B). This configuration provides CLIP with zero-shot capabilities, allowing it potentially to classify images in categories for which it was not explicitly trained, based solely on textual descriptions of those categories (Fig 10C).

For instance, during training, CLIP might see concepts such as “pancreas” and “T1-weighted MR image,” along with various images representing these. However, it may not encounter a combined concept such as “T1-weighted MR image of the pancreas.” Despite this, the zero-shot capability of CLIP allows it to predict new concepts by leveraging its understanding of the relationships and similarities between learned concepts. While

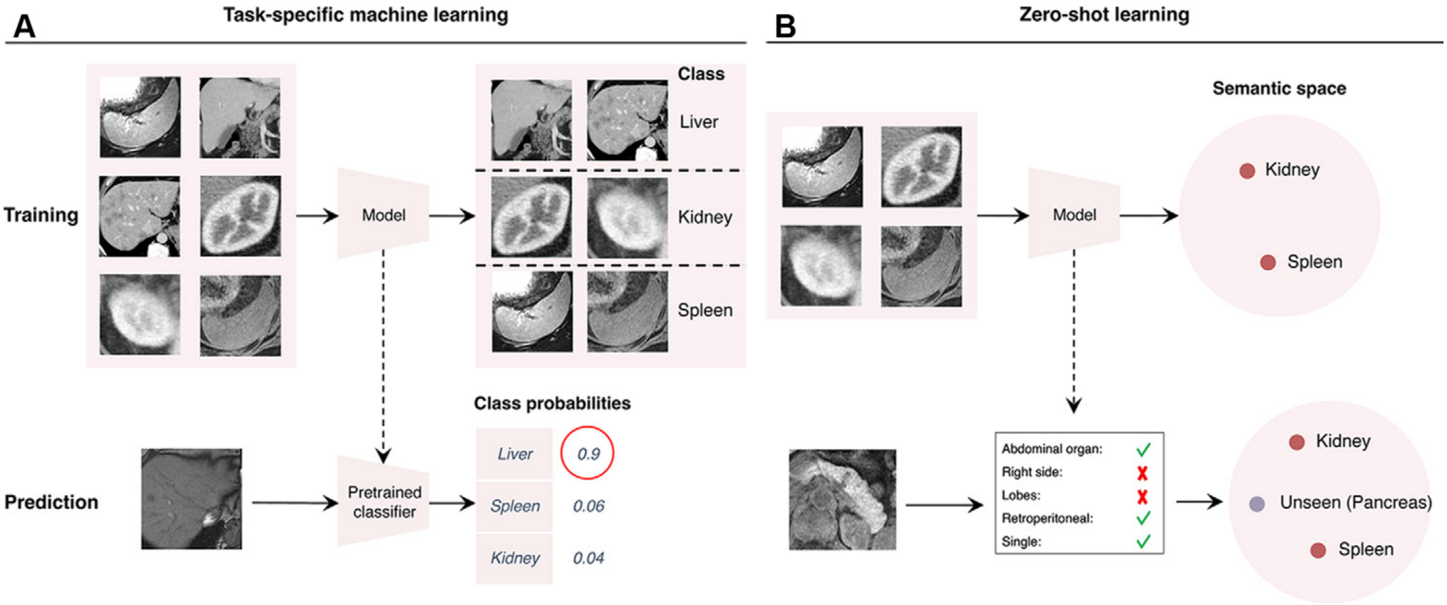


Figure 7. Illustrations show differences between task-specific machine learning and zero-shot learning. **(A)** Task-specific machine learning requires a large, labeled training dataset with all relevant classes to accurately classify new examples during prediction. For accurate prediction, the input image during the prediction phase must be a variation of the classes seen during training, such as different images of the liver in this case. **(B)** Zero-shot learning overcomes the need for labeled examples of unseen classes by enabling the model to generalize from what it has already learned. In this example, the model is trained on a dataset with various radiologic images. During the prediction phase, when it encounters an unseen image (eg, one of the pancreas), the model classifies it based on prior knowledge. It does not specifically recognize or label the image as "pancreas"; instead, it identifies that the image does not match any of the known classes (eg, liver or kidney). Essentially, it performs outlier detection, and it is up to the user to determine that the outlier is the pancreas (33).

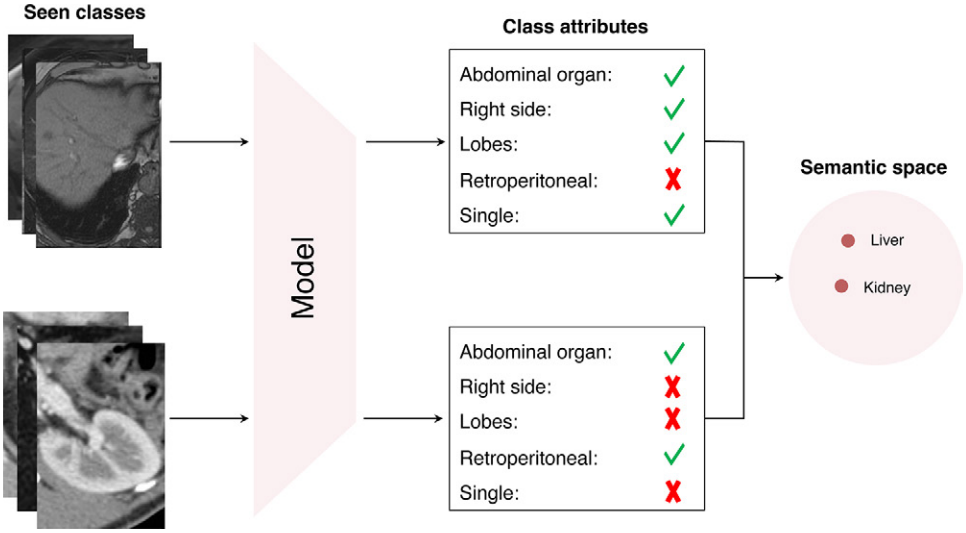


Figure 8. Zero-shot learning. Alongside the labeled data, zero-shot learning requires auxiliary information about the classes, referred to as class (or semantic) attributes. These attributes describe the characteristics or properties associated with each class. For example, the liver (upper images) is an abdominal, single, right-sided lobulated organ without retroperitoneal localization. By generating attributes for each set of images, semantic embeddings are created by mapping these attributes to an embedding space. During training, the model learns to associate the attributes with the class labels using labeled data (35).

CLIP might not have explicitly learned "T1-weighted MR image of the pancreas," it can use its knowledge of individual components to infer that an image shows these elements, even if it has never seen that exact combination before.

Conversely, if "pancreas" is an entirely new concept that CLIP has never encountered, it may not identify the image as a pancreas but can still recognize that it is not a liver or a kidney based on its learned associations. This is why "pancreas" must be a concept that CLIP encountered during training.

Applications.—Contrastive learning from paired images and text is appealing in radiology (17–20,37), given the scarcity of expert-labeled data; this type of learning can leverage domain-specific knowledge found in the association of radiology images with reports. This has prompted an emerging literature on CLIP models specialized for medical image domains (17–20,37). However, most of these developments are still nascent, focusing mostly on chest radiography applications (17–20,37).

Although CLIP is a versatile model, there are several areas where it has constraints and potential drawbacks. While CLIP has zero-shot capabilities, its performance is often lower than that of task-specific models when ample task-specific training data are available (39). The model's performance can also

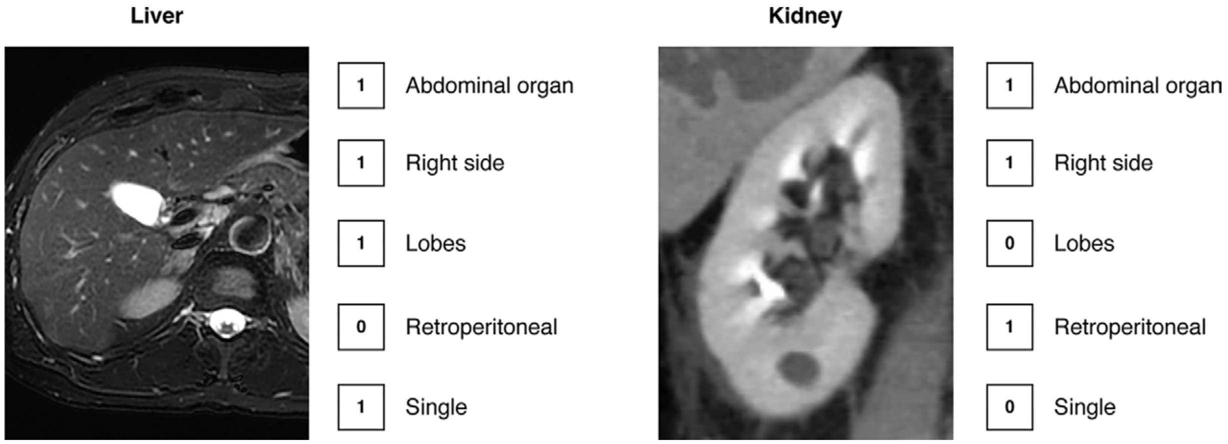
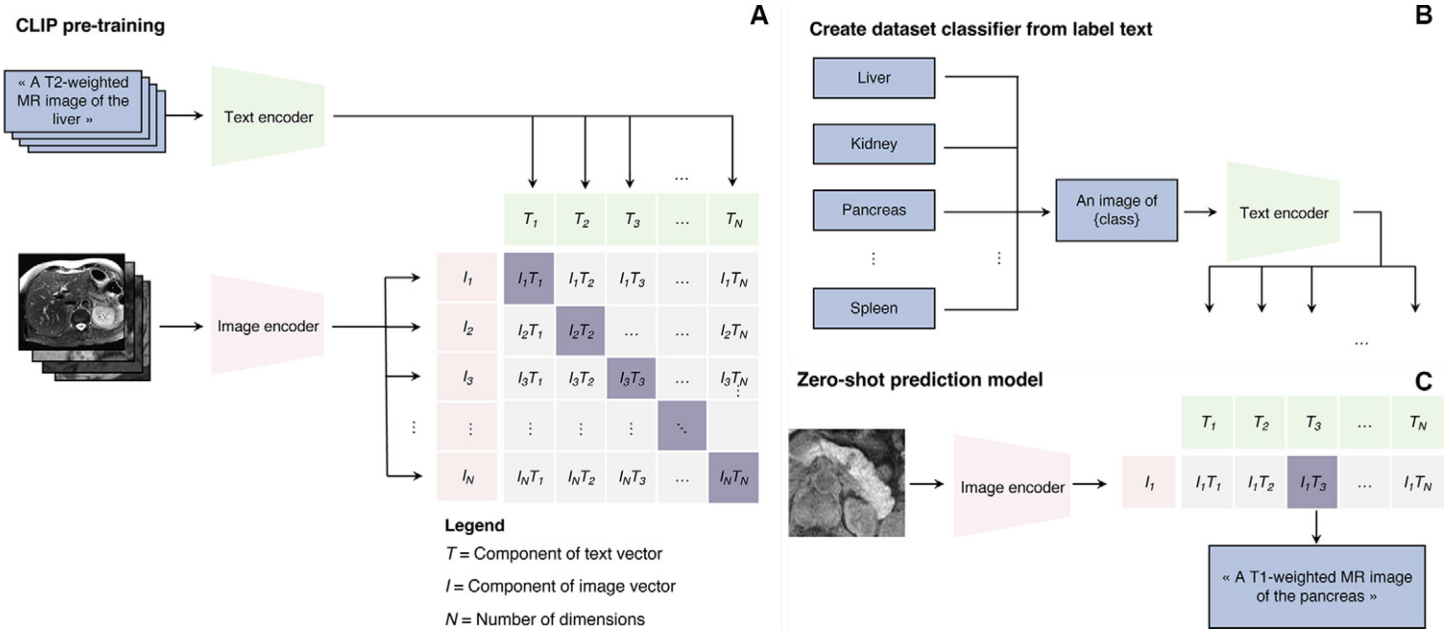


Figure 9. Illustration of semantic embedding. Semantic embedding is a way to represent data (ie, images or words) that captures their meaning. A special set of numerical values is assigned to each example based on its features and relationships with other examples. Examples that have similar characteristics have similar numbers, and examples with different characteristics have different numbers. For instance, both the liver and the kidney shown are right-sided organs (value of 1 for each image), but they differ in their localization (ie, retroperitoneal; divergent values of 1 and 0). By using semantic embedding, models can compare and analyze the content of different examples, even if they have never seen those specific examples before. This is typically achieved through deep learning techniques, where neural networks are trained on large datasets to learn to extract meaningful features that are then transformed into vectors that capture the information in a compact representation (33).


A
CLIP pre-training
« A T2-weighted MR image of the liver »
Text encoder
Image encoder
T_1 T_2 T_3 ... T_N
I_1 I_1T_1 I_1T_2 I_1T_3 ... I_1T_N
I_2 I_2T_1 I_2T_2 ... ... I_2T_N
I_3 I_3T_1 I_3T_2 I_3T_3 ... I_3T_N
⋮ ⋮ ⋮ ⋮ ⋱
I_N I_NT_1 I_NT_2 I_NT_3 ... I_NT_N
B
Create dataset classifier from label text
Liver
Kidney
Pancreas
⋮
Spleen
An image of {class}
Text encoder
C
Zero-shot prediction model
Image encoder
T_1 T_2 T_3 ... T_N
I_1 I_1T_1 I_1T_2 I_1T_3 ... I_1T_N
« A T1-weighted MR image of the pancreas »

Legend
T = Component of text vector
I = Component of image vector
N = Number of dimensions

Figure 10. Illustrations show contrastive language-image pretraining (CLIP). CLIP is a text-image alignment model that pairs embeddings of text and images. **(A)** The architecture of CLIP features a text encoder and an image encoder that produce embeddings in a common space. The training process aligns embeddings of matching image-text pairs while separating those of nonmatching pairs. **(B)** After training, CLIP can classify images based on text descriptions. It converts class descriptions into embeddings with its text encoder and compares them to the image embeddings generated by its image encoder. The image is assigned to the class with the highest similarity score. **(C)** This setup gives CLIP zero-shot capabilities, enabling it to classify images into new categories based only on textual descriptions. For example, to identify axial images of the pancreas, CLIP can use the class description "an axial image of the pancreas" even if it wasn't specifically trained for that category (39).

decrease substantially with shifts in data distribution, where the test data differ greatly from the training data. This limits its reliability in real-world applications where such shifts are expected. To improve its performance in a medical setting, versions of the CLIP model for medical domains have been trained on images and captions extracted from PubMed (40,41).

Image-to-Text Models

Description.—Image-to-text models are generative models designed to generate textual descriptions from images. These models take an image as input and create text descriptions as output.

Key Concepts.—Image-to-text models have two key components: an image encoder and a text decoder. The image encoder first converts the image input into an image embedding. The image encoder of most current text-to-image models is based on transformers. The text decoder uses the image encoder output (ie, the compact image representation) to produce text descriptions of the input image. The text decoder is also typically based on transformers.

The transformer is a deep learning architecture introduced for natural language processing in 2017 in a now-famous article titled "Attention Is All You Need"(42). Transformer models have been proven to be both effective and efficient in understanding the meaning of words in their larger contexts (42) and are now pervasive in natural language processing applications.

A key idea behind the transformer model is self-attention, a computational mechanism that allows the model to focus on different parts of the input data by weighing their relative importance. Attention in deep learning is motivated by how humans pay visual attention to specific parts of an input. An attention component enables the model to "attend" to different parts of the input data, thereby capturing the relationships among these elements, regardless of their distance (Fig 11A) (43).

The Vision Transformer (ViT; *https://github.com/google-research/vision_transformer*) adapts the transformer architecture to images (Fig 11B) (44). Instead of processing an image as a whole, it divides the image into a sequence of smaller patches. Each patch is converted into an embedding, and self-attention can be understood to facilitate "communication" among these embeddings to capture the spatial and contextual relationships among the patches. Given sufficient training data, ViTs have been found to outperform convolutional neural networks in some image-processing contexts (45). The self-attention mechanism may allow the ViT, compared with convolutional neural networks, to more flexibly integrate the global image context in the analysis of image details.

Applications.—Applications of image-to-text models include image captioning (ie, generation of a textual description for a given image) and optical character recognition (ie, conversion of different types of documents into editable and searchable data) (46). These models have the potential for use in medical imaging to generate preliminary radiology reports from imaging studies (47–49).

Text-to-Image Models

Description.—Text-to-image models are generative models designed to generate synthetic images from text descriptions.

Key Concepts.—Similar to image-to-text models, text-to-image models have two key parts: a text encoder and an image decoder. The text encoder usually employs a transformer architecture to convert the text input into a text embedding. The image decoder uses the text embedding to guide the production of synthetic images.

There are several types of generative text-to-image model architectures, each with its own distinctive approach, although the details are beyond the scope of this article. Examples include DALL-E (OpenAI), Stable Diffusion (*www.stablediffusion.com*), and Imagen (*https://imagen.ai.com*) (50–52). Diffusion models, in particular, have recently gained popularity for image generation. A diffusion model iteratively refines a noisy image into a clear image using the text encoding to guide the denoising process (Fig 12) (53,54).

Applications.—Applications of text-to-image models include computer-aided design (eg, creative prototyping by turning textual concepts into visual representations), art generation, and data augmentation (eg, production of synthetic data to train deep learning models) (55). In medical imaging, these models can be used for dataset creation in research or educational settings (56). One example of dataset creation is RoentGen (*https://stanfordmimi.github.io/RoentGen/*), a text-to-image medical domain-adapted model that can generate synthetic chest radiographs when given text prompts (15, 21).

Challenges include possible poor representation of the text description in the generated images and the potential for biased or unrealistic images (57). Biases in training data may contribute to problems with undesired outputs.

Multimodal Models

Description.—Multimodal models integrate multiple types of data such as images, videos, text, and numbers into a unified framework (Fig 13A, 13B).

Key Concepts.—To be integrated together, all data types share a joint embedding space (58). Combining various data types enables multimodal models to have a more comprehensive understanding of input context, potentially leading to a more nuanced output.

Multimodal models have the potential to achieve higher understanding of the input than unimodal models do. By analogy, radiologists review not only images but also associated electronic health records and laboratory results to refine the differential diagnosis. A multimodal model has the potential to combine keywords from the patient history (eg, "right lower quadrant pain") with the content of US images (eg, "adnexal mass") and laboratory results (eg, "elevated beta HCG" [human chorionic gonadotropin level]) to produce a prediction (eg, "ectopic pregnancy").

The architecture of multimodal models extends previously explained concepts (59). At a high level, each type of data (image and text) is processed separately using a specialized encoder to create embeddings. The embeddings from the encoders are combined using a unified multimodal encoder (58). This step aligns the different data embeddings into a joint multimodal representation. A decoder then uses this integrated multimodal embedding to generate the final output, typically text (60).

Applications.—Recent development models connecting images and text have primarily focused on models that take both image and text inputs to generate text outputs. OpenAI recently released the generative pretrained transformer (GPT)-4 Vision (GPT-4 V; *https://openai.com/index/gpt-4v-system-card/*)

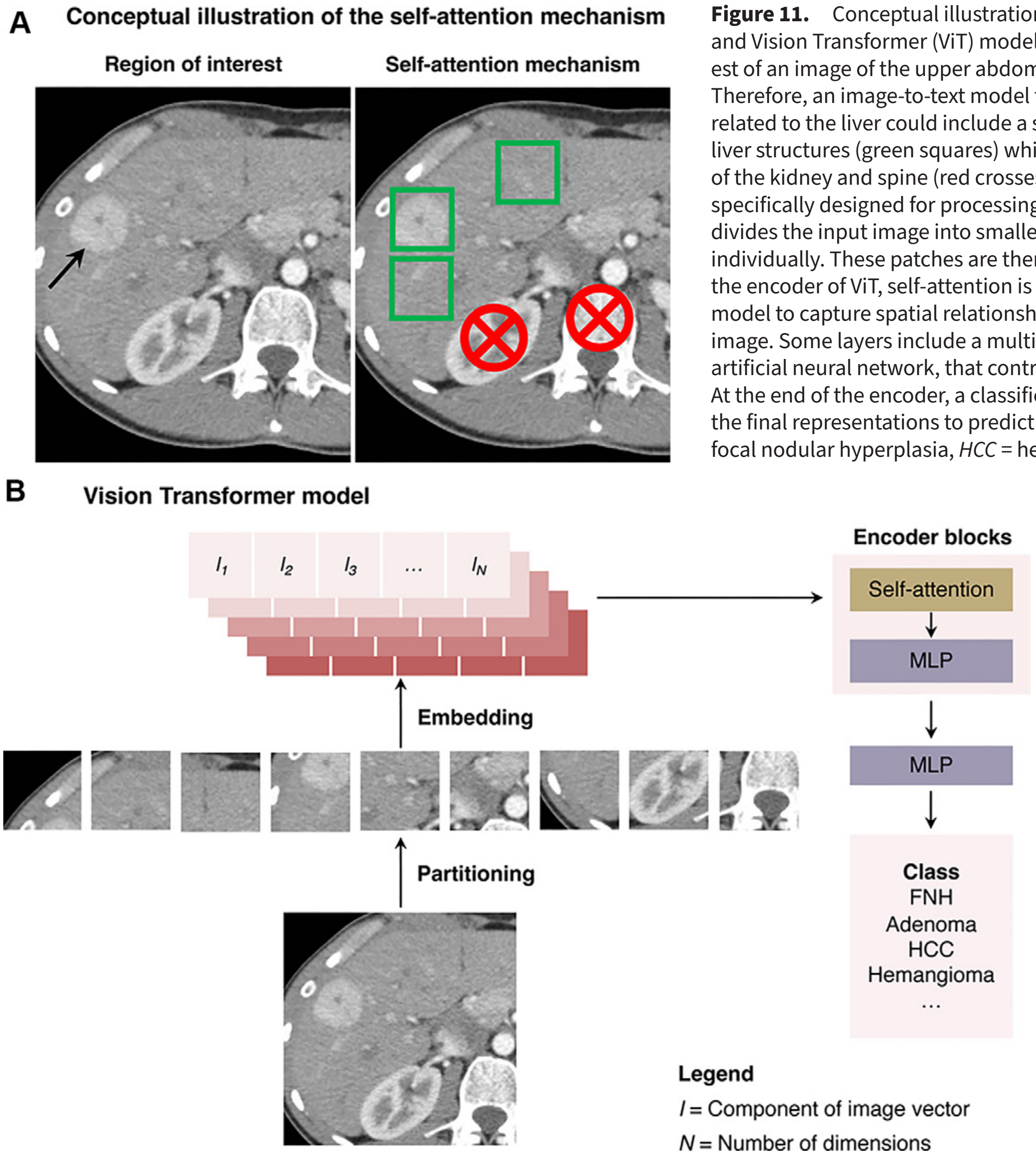


Figure 11. Conceptual illustration of the self-attention mechanism and Vision Transformer (ViT) model. **(A)** For example, the main interest of an image of the upper abdomen could be a liver mass (arrow). Therefore, an image-to-text model trained to describe findings related to the liver could include a self-attention layer focusing on liver structures (green squares) while downweighting the importance of the kidney and spine (red crosses). **(B)** ViT is a transformer model specifically designed for processing visual data, such as images. It divides the input image into smaller patches and processes them individually. These patches are then transformed into embeddings. In the encoder of ViT, self-attention is a key component that allows the model to capture spatial relationships and global context within the image. Some layers include a multilayer perceptron *(MLP)*, a type of artificial neural network, that contributes to deeper feature learning. At the end of the encoder, a classification layer is added, which takes the final representations to predict the image's content (43,44). *FNH* = focal nodular hyperplasia, *HCC* = hepatocellular carcinoma.

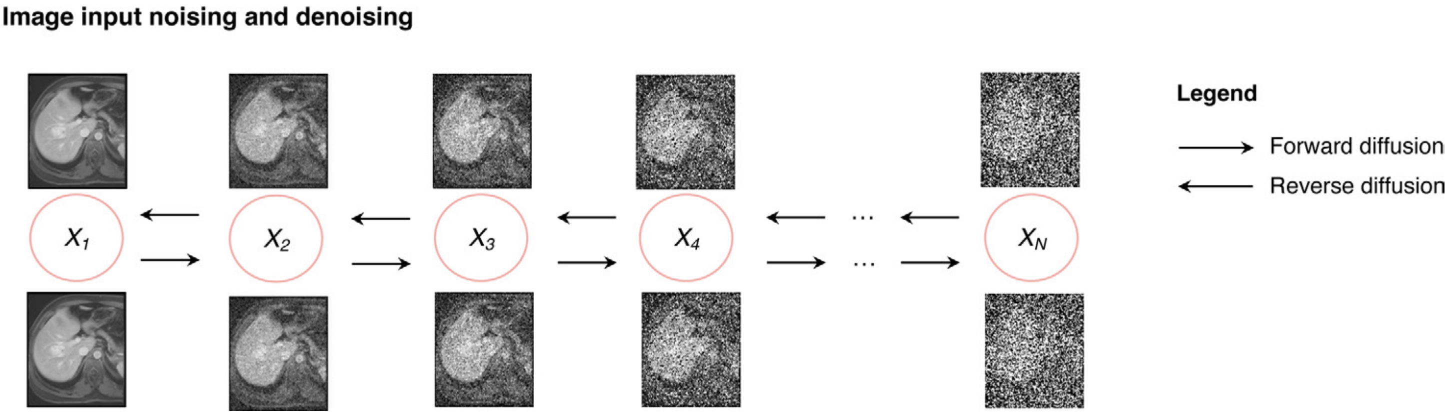


Figure 12. Illustration shows diffusion model architecture. The diffusion model is a type of generative model. It starts with a clear image and gradually adds noise until the image becomes unrecognizable through a forward diffusion process. It then learns to reverse this process by starting with the noisy image and progressively removing the noise to recreate a clear image through a reverse diffusion process. From a technical standpoint, the model generates a series *(N)* of increasingly noisy images by adding Gaussian (random) noise and then reverses this process to produce a series *(N)* of less noisy images, ultimately recovering the original image (54).

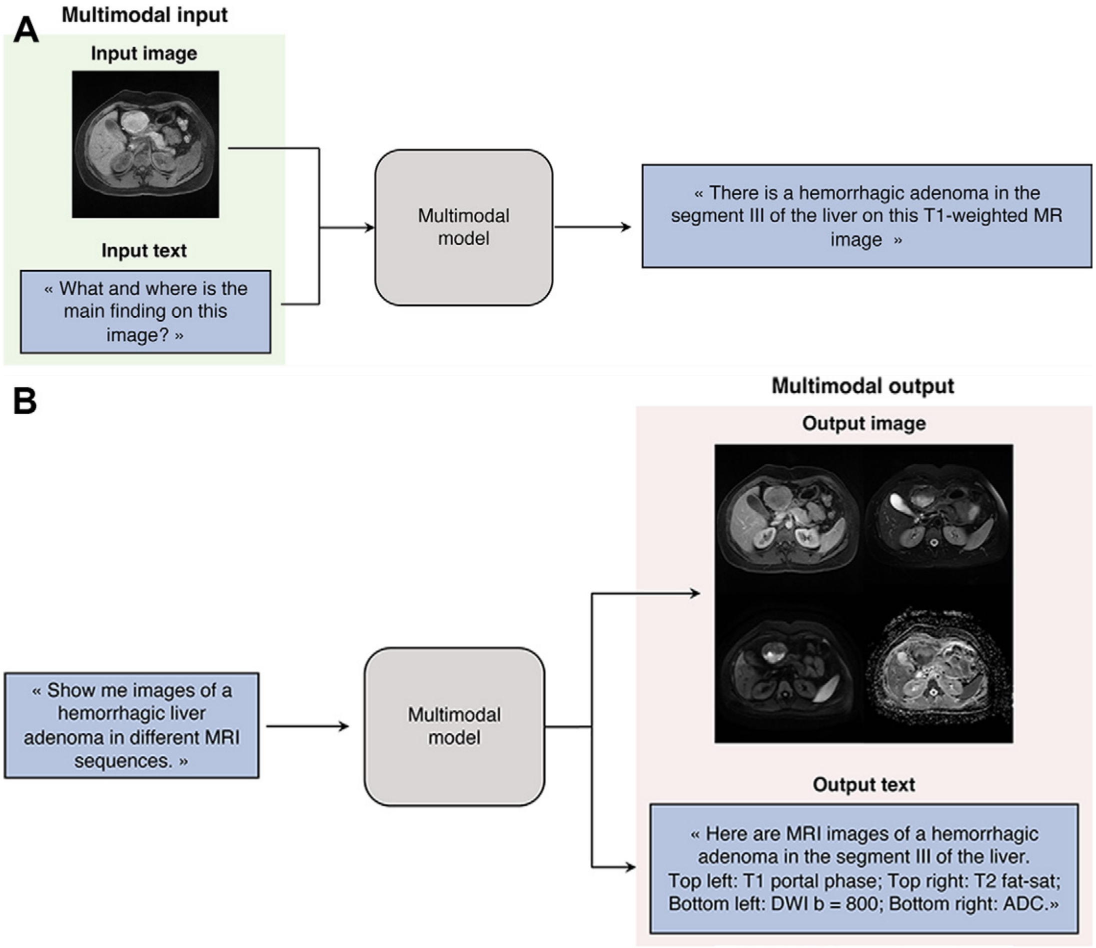

A
Multimodal input
Input image
Input text
« What and where is the main finding on this image? »
Multimodal model
« There is a hemorrhagic adenoma in the segment III of the liver on this T1-weighted MR image »
B
« Show me images of a hemorrhagic liver adenoma in different MRI sequences. »
Multimodal model
Multimodal output
Output image
Output text
« Here are MRI images of a hemorrhagic adenoma in the segment III of the liver. Top left: T1 portal phase; Top right: T2 fat-sat; Bottom left: DWI b = 800; Bottom right: ADC.»

Figure 13. Illustrations show multimodal models. In multimodal models, the input, the output, or both the inputs and output are multimodal. This means that the model can process both image and text data **(A)**, generate both image and text data **(B)**, or process and generate both image and text data (not shown). Images and text are used to show this multimodality, but these models can extend to other types of data, such as video, audio, frequency, and sequencing data. *ADC* = apparent diffusion coefficient, *DWI* = diffusion-weighted imaging, *fat-sat* = fat-saturated, *T1* = T1-weighted, *T2* = T2-weighted.

chatbot, which allows the user to converse with the model using either text or uploaded images. There are also a few models in the open-source space, such as Large Language and Vision Assistant (LLaVA; *https://github.com/haotian-liu/LLaVA*) (61), Bootstrapping Language-Image Pretraining (BLIP; *https://github.com/salesforce/BLIP*) (62), and Flamingo (*https://github.com/mlfoundations/open_flamingo*) (63). A biomedical version is available for each of these models, including LLaVA-Med (*https://github.com/microsoft/LLaVA-Med*) (64), MedBLIP (*https://github.com/qybc/medblip*) (65), and Med-Flamingo (*https://github.com/snap-stanford/med-flamingo*) (66).

By integrating information from various data sources, multimodal models can act as general-purpose assistants, which can be used in multiple domains such as robotics, autonomous driving, e-commerce, and health care (Fig 14) (67). Potential applications in health care include virtual health assistance and clinical decision support, where diverse patient data sources such as medical images, electronic health records, and laboratory results are analyzed to provide more comprehensive and personalized insights (68). The ultimate vision of these models is to augment health care by providing a comprehensive and versatile understanding of medical information across various domains of medicine (15).

Most models linking images and text are trained on internet data, making them successful in general domains. However, models must be fine-tuned with medical data to be suitable for biomedical applications (Fig 15). Domain-specific training by fine-tuning these models on smaller medical datasets is necessary to make them more relevant to radiologic applications. Although data availability hinders the development of these models, authors of some studies (40,41) have pretrained these models on biomedical text and image data.

Conclusion

The integration of both image and text data is an area of recent innovation in AI. Deep learning models connecting images and text data may have great potential to assist radiologists as they face ever-growing volumes of imaging studies and ancillary data. The integration of imaging data with textual clinical information may facilitate automated captioning of medical images, generation of preliminary radiology reports,

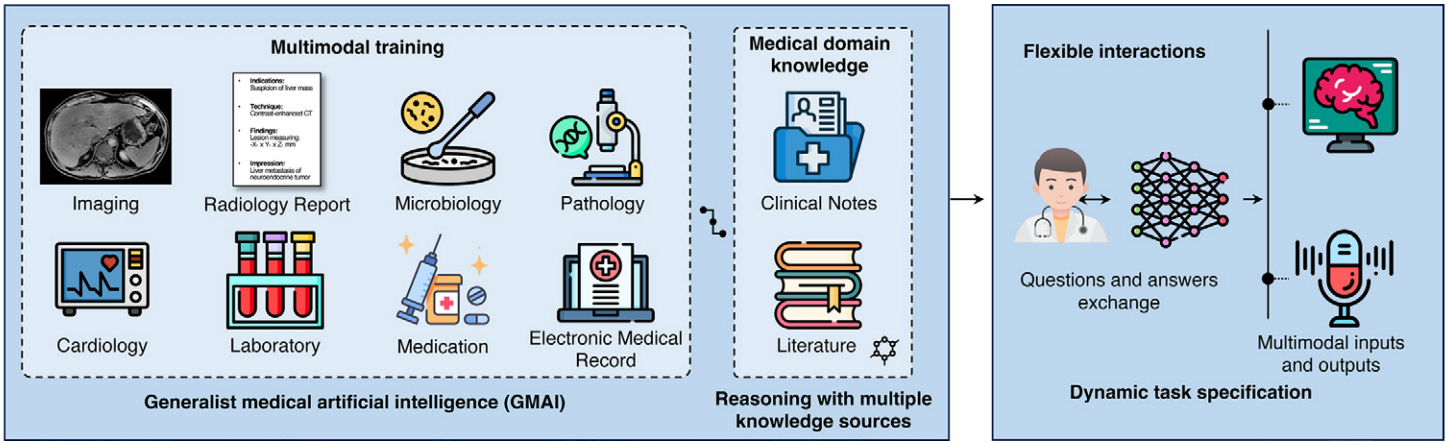

Multimodal training
Imaging
Radiology Report
Microbiology
Pathology
Cardiology
Laboratory
Medication
Electronic Medical Record
Generalist medical artificial intelligence (GMAI)
Medical domain knowledge
Clinical Notes
Literature
Reasoning with multiple knowledge sources
Flexible interactions
Questions and answers exchange
Multimodal inputs and outputs
Dynamic task specification

Figure 14. Illustration shows an overview of a generalist medicine AI *(GMAI)* model pipeline. A GMAI model is trained through techniques such as self-supervised learning on various types of medical data such as imaging, laboratory results, and electronic medical records. To perform clinical reasoning, the model accesses a variety of available medical knowledge including the medical literature and clinical notes. The model can then generate outputs based on inputs that the physician provides (67). (Icons made by surang, Freepik, max.icons, RaftelDesign, and Smashicons from www.flaticon.com.)

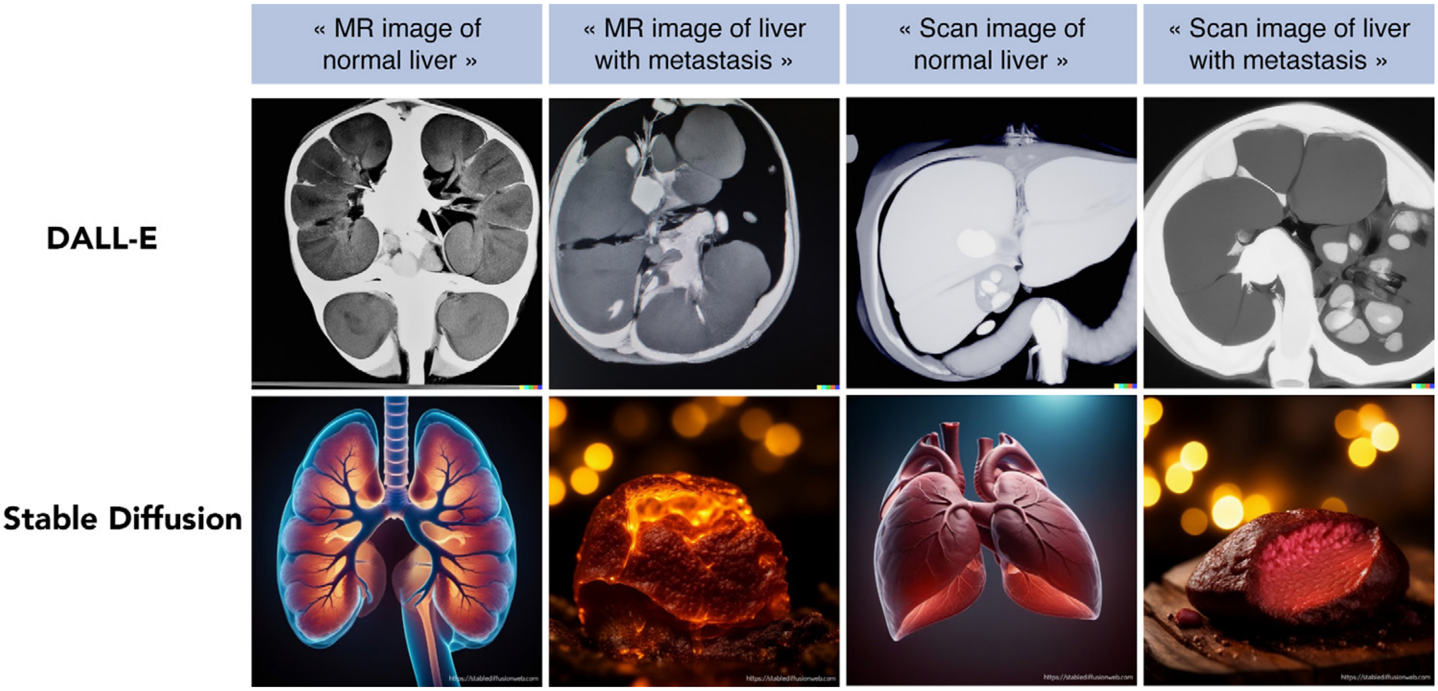

« MR image of normal liver »
« MR image of liver with metastasis »
« Scan image of normal liver »
« Scan image of liver with metastasis »
DALL-E
Stable Diffusion

Figure 15. Examples of synthetic images generated by two text-to-image models, DALL-E 2 (top row) and Stable Diffusion (bottom row), given simple prompts related to medical imaging (50,51). Since these models are trained on large datasets of text and image pairs from the internet, they cannot generalize well to radiology because of the scarcity of openly available medical imaging data, hence the departure from anatomic reality. Future directions include the creation of annotated datasets combining radiology reports and medical images as well as training of these models on these datasets.

and creation of educational images. These advances may enable applications in prioritization of cases, streamlining of clinical workflows, and improvement in diagnostic accuracy. Recognizing the concepts underlying these models can help radiologists in adapting to these new tools and techniques in the future.

Author affiliations.—From the Departments of Radiology, Radiation Oncology, and Nuclear Medicine, Centre hospitalier de l'Université de Montréal, Université de Montréal, 1000 rue Saint-Denis, D03.5431, Montreal, QC, Canada H2X 0C1 (A.N.W., A.C.C., L.L.G., A.T.); Centre de recherche du Centre hospitalier de l'Université de Montréal, Montreal, Quebec, Canada (A.N.W., M.K., L.L.G., E.M., I.B.A., A.T.); Department of Ophthalmology and Visual Sciences, McGill University, Montreal, Quebec, Canada (M.K.); Department of Radiology, Keck School of Medicine of the University of Southern California, Los Angeles, Calif (P.M.C.); Department of Medical Imaging, CISSS Lanaudière, Université Laval, Joliette, Quebec, Canada (A.C.C.); AFX Medical, Montreal, Quebec, Canada (G.C.); Department of Medical Imaging, Western University, London, Ontario, Canada (J.C.); École de Technologie Supérieure, Montreal, Quebec, Canada (I.B.A.); and Institute of Biomedical Engineering, Université de Montréal, Montreal, Quebec, Canada (A.T.). Presented as an education exhibit at the 2023 RSNA Annual Meeting. Received April 9, 2024; revision requested May 23; revision received September 1; accepted September 5. **Address correspondence to** A.T. (email: *an.tang@umontreal.ca*).

Funding.—**L.L.G.** supported by Fonds de Recherche du Québec en Santé (311203), the Fondation de l'Association des Radiologistes du Québec (299979), and the Foundation of the Radiological Society of North America seed grant (RSD2122). **A.T.** supported by Fonds de recherche du Québec en Santé and the Fondation de l'Association des Radiologistes du Québec (298509).

Disclosures of conflicts of interest.—**G.C.** Stock, stock options and other financial and nonfinancial interests with AFX Medical. **J.C.** Chair of the AI Standing Committee, Canadian Association of Radiologists. **I.B.A.** Board member of the Medical Imaging with Deep Learning Foundation. **A.T.** Grants from the Institute of Data Valorization and the Canadian Institutes of Health Research, member of the Liver Imaging Reporting and Data System (LI-RADS) steering committee and cochair of the Lexicon Working Group for LI-RADS, equipment loan from Siemens Healthineers. All other authors, the editor, and the reviewers have disclosed no relevant relationships.

References

1. Winder M, Owczarek AJ, Chudek J, Pilch-Kowalczyk J, Baron J. Are We Overdoing It? Changes in Diagnostic Imaging Workload during the Years 2010-2020 including the Impact of the SARS-CoV-2 Pandemic. Healthcare (Basel) 2021;9(11):1557.
2. Rajpurkar P, Irvin J, Ball RL, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. PLoS Med 2018;15(11):e1002686.
3. Taylor AG, Mielke C, Mongan J. Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: A retrospective study. PLoS Med 2018;15(11):e1002697.
4. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 2018;15(11):e1002683.
5. Neri E, de Souza N, Brady A, et al; European Society of Radiology (ESR). What the radiologist should know about artificial intelligence - an ESR white paper. Insights Imaging 2019;10(1):44.
6. Alikhani M, Khalid B, Stone M. Image-text coherence and its implications for multimodal AI. Front Artif Intell 2023;6:1048874.
7. Li M, Jiang Y, Zhang Y, Zhu H. Medical image analysis using deep learning algorithms. Front Public Health 2023;11:1273253.
8. Oren O, Gersh BJ, Bhatt DL. Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints. Lancet Digit Health 2020;2(9):e486–e488.
9. Pang T, Li P, Zhao L. A survey on automatic generation of medical imaging reports based on deep learning. Biomed Eng Online 2023;22(1):48.
10. Rajpurkar P, Lungren MP. The Current and Future State of AI Interpretation of Medical Images. N Engl J Med 2023;388(21):1981–1990.
11. Selivanov A, Rogov OY, Chesakov D, Shelmanov A, Fedulova I, Dylov DV. Medical image captioning via generative pretrained transformers. Sci Rep 2023;13(1):4171.
12. Chen J, Guo H, Yi K, Li B, Elhoseiny M. VisualGPT: Data-efficient Adaptation of Pretrained Language Models for Image Captioning. arXiv 2102.10407 [preprint] https://arxiv.org/abs/2102.10407. Posted February 20, 2021. Updated March 30, 2022. Accessed March 14, 2025.
13. Monshi MMA, Poon J, Chung V. Deep learning in generating radiology reports: A survey. Artif Intell Med 2020;106:101878https://doi.org/10.1016/j.artmed.2020.101878.
14. Blankemeier L, Cohen JP, Kumar A, et al. Merlin: A Vision Language Foundation Model for 3D Computed Tomography. arXiv 2406.06512 [preprint] https://arxiv.org/abs/2406.06512. Posted June 10, 2024. Accessed March 14, 2025.
15. Moor M, Banerjee O, Abad ZSH, et al. Foundation models for generalist medical artificial intelligence. Nature 2023;616(7956):259–265.
16. Ooi KB, Tan GWH, Al-Emran M, et al. The Potential of Generative Artificial Intelligence Across Disciplines: Perspectives and Future Directions. J Comput Inf Syst 2023;1–32
17. Huang SC, Shen L, Lungren MP, Yeung S. GLoRIA: A Multimodal Global-Local Representation Learning Framework for Label-efficient Medical Image Recognition. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: 3922–3931
18. Wang Z, Wu Z, Agarwal D, Sun J. MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. arXiv 2210.10163 [preprint] https://arxiv.org/abs/2210.10163. Posted October 18, 2022. Accessed March 14, 2025.
19. You K, Gu J, Ham J, et al. CXR-CLIP: Toward Large Scale Chest X-ray Language-Image Pre-training. arXiv 2310.13292 [preprint] https://arxiv.org/abs/2310.13292. Posted October 20, 2023. March 14, 2025.
20. Zhang X, Wu C, Zhang Y, Xie W, Wang Y. Knowledge-enhanced visual-language pre-training on chest radiology images. Nat Commun 2023;14(1):4542.
21. Chambon P, Bluethgen C, Delbrouck JB, et al. RoentGen: Vision-Language Foundation Model for Chest X-ray Generation. arXiv 2211.12737 [preprint] https://arxiv.org/abs/2211.12737. Posted November 23, 2022. March 14, 2025.
22. Bhardwaj A, Khanna P, Kumar S. Pragya. Generative Model for NLP Applications based on Component Extraction. Procedia Comput Sci 2020;167:918–931.
23. Antoniadis P. Latent Space in Deep Learning. Baeldung. https://www.baeldung.com/cs/dl-latent-space. Updated March 18, 2024. Accessed March 14, 2025.
24. Google Developers. Step 3: Prepare Your Data https://developers.google.com/machine-learning/guides/text-classification/step-3. Accessed March 14, 2025.
25. Cheng PM, Montagnon E, Yamashita R, et al. Deep Learning: An Update for Radiologists. RadioGraphics 2021;41(5):1427–1445.
26. Chartrand G, Cheng PM, Vorontsov E, et al. Deep Learning: A Primer for Radiologists. RadioGraphics 2017;37(7):2113–2131.
27. Şenel LK, Utlu I, Yücesoy V, Koc A, Cukur T. Semantic structure and interpretability of word embeddings. arXiv 1711.00331 [preprint] https://arxiv.org/abs/1711.00331. Posted November 1, 2017. Updated May 16, 2018. Accessed March 14, 2025.
28. Jaiswal A, Babu AR, Zadeh MZ, Banerjee D, Makedon F. A survey on contrastive self-supervised learning. arXiv 2011.00362 [preprint] https://arxiv.org/abs/2011.00362. Posted October 31, 2020. Updated February 7, 2021. Accessed March 14, 2025.
29. Alzubaidi L, Zhang J, Humaidi AJ, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 2021;8(1):53.
30. Zhuang F, Qi Z, Duan K, et al. A Comprehensive Survey on Transfer Learning. Proc IEEE 2021;109(1):43–76.
31. Finlayson SG, Subbaswamy A, Singh K, et al. The Clinician and Dataset Shift in Artificial Intelligence. N Engl J Med 2021;385(3):283–286.
32. Antonelli M, Reinke A, Bakas S, et al. The Medical Segmentation Decathlon. Nat Commun 2022;13(1):4128.
33. Kundu R. The Essential Guide to Zero-Shot Learning. https://www.v7labs.com/blog/zero-shot-learning-guide. Published January 6, 2022. Accessed March 14, 2025.
34. Xian Y, Schiele B, Akata Z. Zero-Shot Learning—The Good, the Bad and the Ugly. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: 3077–3086
35. Le Cacheux Y, Le Borgne H, Crucianu M. Zero-shot learning with deep neural networks for object recognition. Multi-faceted Deep Learning: Models and Data 2021:127–150.
36. Wang W, Zheng VW, Yu H, Miao C. A survey of zero-shot learning: Settings, methods, and applications. ACM Trans Intell Syst Technol 2019;10(2):1–37.
37. Zhang Y, Jiang H, Miura Y, Manning CD, Langlotz CP. Contrastive learning of medical visual representations from paired images and text. arXiv 2010.00747 [preprint] https://arxiv.org/abs/2010.00747. Posted October 2, 2020. Updated September 19, 2022. Accessed March 14, 2025.
38. Zhao Z, Liu Y, Wu H, et al. CLIP in Medical Imaging: A Comprehensive Survey. arXiv 2312.07353 [preprint] https://arxiv.org/abs/2312.07353. Posted December 12, 2023. Updated August 10, 2024. Accessed March 14, 2025.
39. Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. arXiv 2103.00020 [preprint] https://arxiv.org/abs/2103.00020. Posted February 26, 2021. Accessed March 14, 2025.
40. Lin W, Zhao Z, Zhang X, et al. Pmc-clip: Contrastive language-image pre-training using biomedical documents. arXiv 2303.07240 [preprint] https://arxiv.org/abs/2303.07240. Posted March 13, 2023. Accessed March 14, 2025.
41. Zhang S, Xu Y, Usuyama N, et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv 2303.00915 [preprint] https://arxiv.org/abs/2303.00915. Posted March 2, 2023. Updated January 16, 2024. Accessed March 14, 2025.
42. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. arXiv 1706.03762 [preprint] https://arxiv.org/abs/1706.03762. Posted June 12, 2017. Updated August 2, 2023. Accessed March 14, 2025.
43. Weng L. Attention? Attention! Lil'Log. https://lilianweng.github.io/posts/2018-06-24-attention/. Published June 24, 2018. Accessed March 14, 2025.
44. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 2010.11929 [preprint] https://arxiv.org/abs/2010.11929. Posted October 22, 2020. Updated June 30, 2021. Accessed March 14, 2025.
45. Maurício J, Domingues I, Bernardino J. Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. Appl Sci (Basel) 2023;13(9):5521.
46. He X, Deng L. Deep Learning for Image-to-Text Generation: A Technical Overview. IEEE Signal Process Mag 2017;34(6):109–116.

47. Alfarghaly O, Khaled R, Elkorany A, Helal M, Fahmy A. Automated radiology report generation using conditioned transformers. Informatics in Medicine Unlocked. 2021;24:100557.
48. Hamamci IE, Er S, Menze B. CT2Rep: Automated Radiology Report Generation for 3D Medical Imaging. arXiv 2403.06801 [preprint] https://arxiv.org/abs/2403.06801. Posted March 11, 2024. Updated July 4, 2024. Accessed March 14, 2025.
49. Yang S, Wu X, Ge S, Zhou SK, Xiao L. Knowledge matters: Chest radiology report generation with general and specific knowledge. Med Image Anal 2022;80:102510.
50. Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M. Hierarchical text-conditional image generation with clip latents. arXiv 2204.06125 [preprint] https://arxiv.org/abs/2204.06125. Posted April 13, 2022. Accessed March 14, 2025.
51. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. arXiv 2112.10752 [preprint] https://arxiv.org/abs/2112.10752. Posted December 20, 2021. Updated April 13, 2022. Accessed March 14, 2025.
52. Saharia C, Chan W, Saxena S, et al. Photorealistic text-to-image diffusion models with deep language understanding. arXiv 2205.11487 [preprint] https://arxiv.org/abs/2205.11487. Posted May 23, 2022. Accessed March 14, 2025.
53. Yang L, Zhang Z, Song Y, et al. Diffusion models: A comprehensive survey of methods and applications. arXiv 2209.00796 [preprint] https://arxiv.org/abs/2209.00796. Posted September 2, 2022. Updated June 24, 2024. Accessed March 14, 2025.
54. O'Connor R. How Imagen Actually Works. AssemblyAI. https://www.assemblyai.com/blog/how-imagen-actually-works/. Published June 23, 2022. Accessed March 14, 2025.
55. Elasri M, Elharrouss O, Al-Maadeed S, Tairi H. Image Generation: A Review. Neural Process Lett 2022;54(5):4609–4646.
56. Adams LC, Busch F, Truhn D, Makowski MR, Aerts HJWL, Bressem KK. What Does DALL-E 2 Know About Radiology? J Med Internet Res 2023;25:e43110.
57. Croitoru FA, Hondru V, Ionescu RT, Shah M. Diffusion Models in Vision: A Survey. IEEE Trans Pattern Anal Mach Intell 2023;45(9):10850–10869.
58. Girdhar R, El-Nouby A, Liu Z, et al. ImageBind One Embedding Space to Bind Them All. arXiv 2305.05665 [preprint] https://arxiv.org/abs/2305.05665. Posted May 9, 2023. Updated May 31, 2023. March 14, 2025.
59. Greg Corrado. Google Research. https://research.google/people/gregcorrado/. Accessed March 14, 2025.
60. Koh JY, Fried D, Salakhutdinov RR. Generating images with multimodal language models. arXiv 2305.17216 [preprint] https://arxiv.org/abs/2305.17216. Posted May 26, 2023. Updated October 13, 2023. Accessed March 14, 2025.
61. Liu H, Li C, Wu Q, Lee YJ. Visual instruction tuning. arXiv 2304.08485 [preprint] https://arxiv.org/abs/2304.08485. Posted April 17, 2023. Updated December 11, 2023. Accessed March 14, 2025.
62. Li J, Li D, Xiong C, Hoi S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. arXiv 2201.12086 [preprint] https://arxiv.org/abs/2201.12086. Posted January 28, 2022. Updated February 15, 2022. Accessed March 14, 2025.
63. Alayrac JB, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning. arXiv 2204.14198 [preprint] https://arxiv.org/abs/2204.14198. Posted April 29, 2022. Updated November 15, 2022. Accessed March 14, 2025.
64. Li C, Wong C, Zhang S, et al. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. arXiv 2306.00890 [preprint] https://arxiv.org/abs/2306.00890. Posted June 1, 2023. Accessed March 14, 2025.
65. Chen Q, Hu X, Wang Z, Hong Y. MedBLIP: Bootstrapping language-image pre-training from 3d medical images and texts. arXiv 2305.10799 [preprint] https://arxiv.org/abs/2305.10799. Posted May 18, 2023. Accessed March 14, 2025.
66. Moor M, Huang Q, Wu S, et al, eds. Med-flamingo: a multimodal medical few-shot learner. arXiv 2307.15189 [preprint] https://arxiv.org/abs/2307.15189. Posted July 27, 2023. Accessed March 14, 2025.
67. Wu J, Gan W, Chen Z, Wan S, Yu PS. Multimodal Large Language Models: A Survey. arXiv 2311.13165 [preprint] https://arxiv.org/abs/2311.13165. Posted November 22, 2023. Accessed March 14, 2025.
68. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. Nat Med 2022;28(9):1773–1784.