electronics

MDPI

Review

Recent Advances in Synthesis and Interaction of Speech, Text, and Vision

Laura Orynbay [1,*], Bibigul Razakhova [1], Peter Peer [2], Blaž Meden [2] and Žiga Emeršič [2]

[1] Department of Artificial Intelligence Technologies, L.N. Gumilyov Eurasian National University, Astana 010008, Kazakhstan; bibigul.razakhova@gmail.com

[2] Faculty of Computer and Information Science, University of Ljubljana, SI-1000 Ljubljana, Slovenia; peter.peer@fri.uni-lj.si (P.P.); blaz.meden@fri.uni-lj.si (B.M.); ziga.emersic@fri.uni-lj.si (Ž.E.)

* Correspondence: laura.aktobe.kz@gmail.com

Abstract: In recent years, there has been increasing interest in the conversion of images into audio descriptions. This is a field that lies at the intersection of Computer Vision (CV) and Natural Language Processing (NLP), and it involves various tasks, including creating textual descriptions of images and converting them directly into auditory representations. Another aspect of this field is the synthesis of natural speech from text. This has significant potential to improve accessibility, user experience, and the applications of Artificial Intelligence (AI). In this article, we reviewed a wide range of image-to-audio conversion techniques. Various aspects of image captioning, speech synthesis, and direct image-to-speech conversion have been explored, from fundamental encoder–decoder architectures to more advanced methods such as transformers and adversarial learning. Although the focus of this review is on synthesizing audio descriptions from visual data, the reverse task of creating visual content from natural language descriptions is also covered. This study provides a comprehensive overview of the techniques and methodologies used in these fields and highlights the strengths and weaknesses of each approach. The study emphasizes the importance of various datasets, such as MS COCO, LibriTTS, and VizWiz Captions, which play a critical role in training models, evaluating them, promoting inclusivity, and solving real-world problems. The implications for the future suggest the potential of generating more natural and contextualized audio descriptions, whereas direct image-to-speech tasks provide opportunities for intuitive auditory representations of visual content.

Keywords: text-free image; audio description; image captioning; text-to-speech; image-to-speech; text-to-image; synthesis; data generation; Computer Vision; Natural Language Processing; Artificial Intelligence

check for updates

Citation: Orynbay, L.; Razakhova, B.; Peer, P.; Meden, B.; Emeršič, Ž. Recent Advances in Synthesis and Interaction of Speech, Text, and Vision. *Electronics* **2024**, *13*, 1726. https://doi.org/10.3390/electronics13091726

Academic Editor: Cheng He

Received: 12 March 2024
Revised: 8 April 2024
Accepted: 24 April 2024
Published: 30 April 2024

1. Introduction

In the modern era, where visual information is prevalent, accessibility is more important than ever before. Since over two billion people suffer from visual impairment globally, it is vital to ensure equitable access to visual content [1].

Audio description is essential for accessibility, especially for visually impaired people. This is essential for inclusion because it guarantees that the information supplied visually may also be accessed aurally. The following are the main arguments as to why audio description is crucial for accessibility:

- Equal access to information: An audio description guarantees an equal chance for individuals with visual impairments to view visual media such as live events, TV shows, films, and instructional videos.
- Social inclusion: The audio gives visually impaired persons a sense of community and belonging by allowing them to take part in common cultural events and recreational events, including athletic events, art exhibitions, and museum tours.

- Independent navigation: For the blind or visually impaired in public spaces, audio description is a crucial part of independent navigation. This makes it easier for people to walk safely and offers insightful information about their surroundings.
- Accessibility in the learning process: In educational contexts, audio description is a method used to deliver visual elements such as charts and graphs. It is crucial that visually impaired students participate in their studies and be completely incorporated into various topics.
- Employability: People with visual impairments must have access to job descriptions that are easy to read. This promotes inclusion and equal employment opportunities by allowing individuals to interact with visual presentations, diagrams, and other work-related materials.

Converting image data into audio descriptions is useful not only for people with visual impairments but also to improve usability for a wider audience:

- Multitasking and convenience [2]: Audio descriptions provide a convenient alternative for people who may be working in multitasking mode or cannot focus on visual content, allowing them to perceive information by ear while engaged in other activities.
- Language diversity: People who speak different languages or have different levels of proficiency in the language of visual content can benefit from audio descriptions because they provide an oral explanation that overcomes language barriers [3].
- Learning styles: Individuals differ in how they absorb information, and some may gain more or prefer auditory knowledge [4]. Audio descriptions are provided for people who retain information better by listening than by seeing.

Essentially, audio description serves as a bridge between the visually impaired and the outside world. This goes beyond being accessible. This is the basic requirement for building an inclusive society in which all people, regardless of their abilities, can fully participate in all areas of life.

The integration of Computer Vision (CV) and Natural Language Processing (NLP) is required to convert image data into audio descriptions. Through this collaboration between NLP and CV technologies, individuals with visual impairments can fully comprehend visual content and transform it into audio descriptions. One of the main advantages of this integration is the intelligent interpretation of data. NLP provides a linguistic context for processing visual content in a CV, which allows for a deeper and more advanced understanding of images.

The article provides a comprehensive, up-to-date, and insightful study of the current state and future directions of Artificial Intelligence (AI) technologies related to image captioning, Text-to-Speech (TTS), and new Image-to-Speech and Text-to-Image areas, as shown in the diagram in Figure 1. The study discusses different datasets specific to indirect and direct Image-to-Speech tasks, emphasizing their unique characteristics and applications. The article recognizes the role of datasets in model training and evaluation and highlights their importance in various aspects of research. It provides valuable information for researchers, practitioners, and developers in these fields.

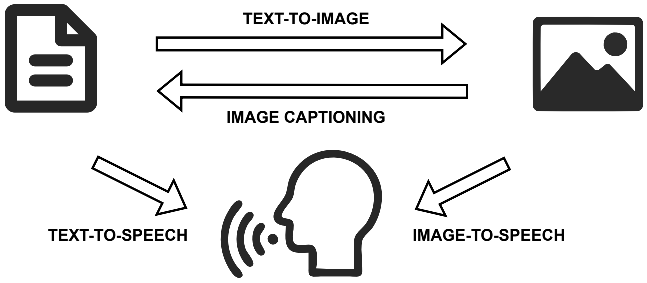

Figure 1. Diagram showing relations between tasks of: image captioning, text-to-speech, image-to-speech, and text-to-image.

The first section of this article explains the symbiotic relationship between NLP and CV in the context of converting visual data into meaningful audio descriptions and provides examples of successful applications of AI-based virtual assistance. In the second section, a structured taxonomy of methods is developed, categorizing them based on their ability to handle different types of data, such as vision, text, and sound. The first subsection begins with image captioning, a fundamental task in AI, where visual understanding intersects with linguistic expressions. Next, we explore Text-to-Speech technologies, following their evolution from traditional synthesis methods to the emergence of neural TTS based on deep learning techniques. We also discuss recent studies on Image-to-Speech systems that enable computers to produce spoken descriptions from images without the need for accompanying text. In the fourth subsection, we examine models in the developing field of Text-to-Image conversion that synthesize visual content from textual descriptions. Finally, the third section explores and compares datasets critical to three vital areas of AI: image captioning, Text-to-Speech synthesis, and Image-to-Speech tasks.

2. Integration of Natural Language Processing (NLP) and Computer Vision (CV)

2.1. Explanation of the Symbiotic Relationship between NLP and CV

The rapid development of deep learning algorithms is one of the main drivers of the convergence of language and visual processing in the modern era [5,6]. Advances in deep learning have raised the bar for both Computer Vision (CV) and Natural Language Processing (NLP), with each area showing impressive growth in different tasks [7]. CV has advanced significantly in object detection, semantic segmentation, and visual-content classification. NLP has also seen growing interest, especially in large-scale unlabeled corpora used for unsupervised pre-training of language models.

To address complex tasks, there is currently growing interest in integrating linguistic and visual information, thus bridging traditionally independent domains. Approaches to this integration task should provide a deep understanding of the textual or visual content. These methods must translate textual content using visual cues to distinguish it, identify objects and reason relationships, and create grammatically correct descriptions of visual content. Overcoming these obstacles could lead to practical applications such as assisting blind people; automatic monitoring; self-driving cars; facilitating human–computer interactions; city navigation; and providing a thorough testing ground for CV and NLP systems [7].

2.2. Importance of Joint Processing in Converting Visual Data to Meaningful Audio Descriptions

The conversion of visual data into coherent and meaningful audio descriptions relies heavily on collaborative processing, particularly the joint integration of NLP and CV technologies. Audio description is a means for people with visual impairments to access the visual elements of various types of activities, including theatre, media, and visual arts, where imagery plays an important role. Using concise, vivid, and imaginative language, audio descriptors ensure that visual information is accessible to segments of the population to which it may be inaccessible or only partially accessible [8,9]. The purpose of the audio description technique is to clearly and concisely convey important visual aspects such as actions, surroundings, gestures, facial expressions, and other details that contribute to a deeper understanding of the situation [10]. This goes beyond mere narration.

Hansjorg Bittner [11] noted that the visual elements covered by audio description include:

- Form: Includes characters, places, words, or any recognizable shape or object.
- Motion: Refers to any state or sign of motion, including actions and the passage of time.
- color: Includes the hue and skin tone of the characters.
- Sound: Refers to sounds that can only be discerned through visual cues.
- Camera Perspective: Includes aspects such as point of view, scale, bird's eye view, and camera special effects.

- Supporting Information: Consists of extraneous information and details such as changes in information.

The equal-access principle is the foundation of audio description. This is predicated on the fundamental idea that everyone ought to have equal access to visual content perception and understanding regardless of their level of visual ability. To ensure that individuals with visual impairments can participate in the visual narrative and experience emotions, subtleties, and nuances provided by visual content, audio description aims to improve the non-visual experience rather than replace the visual experience [12].

2.3. Examples of Successful AI-Powered Visual Assistance Applications

The effective integration of Artificial Intelligence (AI) in transforming visual data into meaningful auditory descriptions has been demonstrated by several successful examples that particularly help people with visual impairment. Consider the various functionalities and attributes inherent in these tools and systematically categorize them for ease of understanding.

2.3.1. Object Recognition and Text-to-Speech

- Seeing AI (Microsoft) [13], a free app designed for blind and visually impaired people, uses Artificial Intelligence (AI) to describe the surroundings audibly. Its features include instant text voicing, document text recognition, barcode scanning, facial recognition with age and gender estimation, currency recognition, scene description, audio-augmented reality for space exploration, indoor navigation, color identification, handwriting reading, light estimation, and integration with other image recognition applications. This multifaceted tool allows users to easily navigate their surroundings.
- Envision AI [14], an award-winning Optical Character Recognition (OCR) app designed for the visually impaired, uses AI and OCR to audibly interpret the visual world, promoting independence. With full spoken language support, it quickly reads a text in 60 languages, scans documents, recognizes PDFs and images, interprets handwritten notes, and describes the scenes. The app also detects colors, scans barcodes for product information, and recognizes nearby people and objects. Envision allows the sharing of images and documents from different applications and provides voice descriptions to enhance accessibility.
- TapTapSee [15] is a specialized application designed to assist blind or visually impaired people in identifying objects during their daily activities. Users can take pictures by tapping on any part of the screen, which makes it easier to photograph two-dimensional or three-dimensional objects from different angles. The app then provides voice identification depending on VoiceOver activation. Recognized for its usefulness, TapTapSee has received notable awards, including the American Foundation for the Blind 2014 Access Award Recipient and RNIB (Royal National Institute of Blind People) recognition as an App of the Month in March 2013. The application includes features such as image recognition; the ability to repeat the last identification; uploading; saving images from a photographic film with appropriate definitions; and sharing the results via text messages, email, or social networks. Developed by CloudSight Inc., a Los Angeles-based technology company specializing in image captioning and understanding, TapTapSee is a sophisticated solution that promotes the independence of the visually impaired.

Additionally, applications like Aipoly Vision [16] and iDentifi [17], which are not currently listed in application markers, were discovered on the Internet.

2.3.2. Navigation and Location Assistance

- BlindSquare [18], an innovative navigation solution for people with visual impairments, combines GPS, compass, and FourSquare data to provide comprehensive assistance indoors and outdoors. Developed in collaboration with insights from the blind community, the app uses algorithms to extract relevant information transmitted through high-quality speech synthesis. Enabling voice commands as a premium service increases user control. BlindSquare is a universal GPS solution, offering step-by-step instructions and searching for detailed information regarding nearby locations and compatibility with other navigation applications. Acting as a four-square client, the application supports registration and related actions. It is a paid application; it supports 25 languages; and it has received awards such as the GSMA (Groupe Speciale Mobile Association) Global Mobile Awards 2013 for the best mobile product or service in the field of healthcare, BlindSquare.

- Aira [19] is a visual translation placement service that provides real-time communication between visually impaired people and professionally trained visual translators through the Aira Explorer application. The application is accessed by pressing a button on the main screen; assisting with requests 24/7; and increasing independence and efficiency in describing, reading, explaining, and navigating in various environments. Live video streaming includes GPS location detection, which allows agents to interact with the user's environment through an integrated dashboard that includes web data, maps, location tracking, search engines, text messages, and rideshare integration. Aira Access distributes the service to organizations that are members of the Aira Access network, allowing visually impaired people to use the service for free on partner sites, contributing to accessibility and inclusivity.

2.3.3. Face Recognition and Identification

NoorCam MyEye [20] is an advanced wearable voice-activated AI assistive technology designed for various levels of vision loss. Designed to be unobtrusive and easy to carry, the device works offline, reducing data privacy concerns and providing adaptability to different environments. The high-precision laser guidance used by an intelligent camera allows for the transmission of visual information in real-time, including text reading, face recognition, product identification using barcode scanning, checking paper money denominations, and surface color recognition. This comprehensive solution is a portable and effective means of helping people with visual impairments by offering instant and accurate information thanks to innovative AI functions.

2.3.4. Assistance from Sighted Volunteers

Be My Eyes app [21] is a comprehensive tool that combines three different functions to help people with visual impairment. Used by more than half a million people worldwide, the app connects with an extensive network of volunteers, who can provide visual descriptions in 185 languages. Winning awards, such as the 2021 Apple Design Award for Best Social Impact App and inclusion in Time magazine's 2023 Best Inventions list, Be My Eyes demonstrates a significant contribution to inclusion and empowerment on a global scale.

2.3.5. General Visual Assistance

Lookout (by Google) [22], an application for auxiliary vision, uses CV to improve the efficiency of people with low vision or blindness. Developed in consultation with the blind and visually impaired community, the app is in line with Google's commitment to making information universally accessible. Offering six modes, including the newly introduced image mode for detailed image description and question-based interaction (in English only in the US, UK, and Canada), Lookout facilitates tasks such as text viewing and presentation by ear (text mode), quick identification of packaged products using labels and barcodes (food label mode), capturing entire pages of text or handwriting (document mode), the quick identification of banknotes (currency mode supporting US dollars, Euro and Indian Rupees), and a research mode (beta version) providing information about the surroundings. Supporting more than 20 languages, Lookout is compatible with Android devices running on Android 6 and above, with a recommended RAM of 2 GB or more.

2.3.6. Wearable Devices

CyberEyez [23] serves as a wearable magnification solution specifically designed for people with low vision, covering conditions such as macular degeneration, retinitis pigmentosa, nystagmus, stroke, and traumatic brain injury (TBI). This technology uses cost-effective and easily accessible equipment and smartphones, offering an affordable and adaptable solution to achieve various life goals in educational, professional, and everyday contexts. The system combines a virtual reality headset with Bluetooth joystick remote control, facilitating activities such as zooming, gaming, and virtual reality experiences. The implementation of this innovative solution is characterized by its cost-effectiveness: options such as Google Cardboard headsets are paid.

Eyesynth's NIIRA smart glasses [24] represent an innovative solution for individuals with blindness and poor vision. These glasses use 3D technology to enable users to identify shapes, measure depths, and locate objects precisely. They employ bone conduction audio and transmit sound through the skull to free the ears to hear the environment. With a battery life of up to 10 h, the NIIRA ensures continuous operation. It offers two modes: tracking and panoramic, allowing intuitive perception of surroundings. NIIRA adapts to individual needs and provides solutions for both partial and complete vision loss. It incorporates real-time audio processing, which operates even in complete darkness.

eSight glasses [25] are innovative wearable devices designed to assist individuals with macular degeneration and other visual impairments. The glasses feature a small high-definition camera that captures the user's environment, and advanced algorithms optimize and enhance footage in real-time. The enhanced image is then displayed on OLED screens, providing users with up to 20/20 vision enhancements. These glasses operate wirelessly and are hands-free, thus allowing users to move freely. With a battery life of up to 3 h and touch controls for features, such as zoom and contrast adjustment, users can easily engage in daily activities indoors and outdoors.

Sight Plus [26] is a wearable low-vision device designed to help individuals with various visual impairments independently perform daily activities. This device is beneficial for patients with central vision loss, such as those affected by age-related macular degeneration (AMD) or Stargardt's disease, as well as for those with conditions that affect the entire visual field, such as albinism or optic neuritis. Unlike traditional reading aids, Sight Plus offers portability and allows users to read comfortably anywhere and at any time. It provides magnification capabilities, enabling users to see faces, enjoy television shows and movies from anywhere in the room, engage in online activities such as viewing whiteboards and screens, and perform work or study tasks. The hands-free design of the device facilitates the enjoyment of hobbies and crafts as well as the playing of musical instruments. Sight Plus supports exploration and travel, making it easier for users to navigate their surroundings and capture images to revisit memories later.

NuEyes Smartglasses [27] provide innovative solutions for people with visual impairments. These compact wearable devices, which weigh only 102 g, offer a powerful solution for enhancing vision and connectivity. The glasses feature a 1080p display that provides a 43-degree field of view. This allows users to stay connected to loved ones without requiring cumbersome equipment. Removable visual prosthetics keep hands free and assist visually impaired individuals in seeing again, enabling them to participate fully in everyday activities. The previously mentioned NoorCam My Eye [20] is a wearable Artificial Intelligence technology that can be activated by voice.

Tailoring multimodal solutions to the diverse needs of visually impaired individuals is crucial. Future research directions may include adaptable output options (e.g., customizable audio description detail, braille output, and haptic feedback) and individually adaptable input methods (e.g., voice commands, gesture-based controls) tailored to individual needs. Nevertheless, the development of these user-centered designs will need to include extensive collaboration with individuals across the spectrum of visual impairment. Only then can we develop solutions with customizable output modalities (e.g., varying levels of audio description verbosity, braille output options) and flexible interaction methods to accommodate user preferences and abilities.

2.3.7. Types of Hardware

Hardware used for AI model training and inference includes a vast array of options. However, mostly Nvidia's CUDA is used, and the hardware can be split into the following two groups:

- GPUs (Graphics Processing Units): GPUs remain the workhorse of most deep learning applications, including generative AI, due to their parallel processing capabilities. Top-of-the-line options include examples such as NVIDIA A100 and H100 Tensor Core GPUs, which are designed specifically for AI and scientific workloads. GPUs used widely by researches for either intermediate or more managable experiments include NVIDIA RTX 30 and 40 series—high-end consumer GPUs with substantial power for generative models.
- Specialized AI Accelerators: These chips offer even greater efficiency and performance for specific AI workloads. One popular option is Google TPUs (Tensor Processing Units), optimized for Google's TensorFlow framework and commonly used in Google Cloud. Graphcore IPUs (Intelligence Processing Units) are designed for flexibility and handling large, complex models. Additionally, AWS Trainium and Inferentia, Amazon's custom AI accelerator chips, may be a viable option at the time of the writing.

With considerations for Text, Image, and Speech, the following requirements need to be taken into consideration:

- Model Size and Complexity: Larger, more complex generative models require more powerful hardware with higher memory capacity.
- Image Tasks: Models for image generation often demand higher GPU memory (VRAM) compared to purely text-based models.
- Speech Tasks: Generating realistic speech can be computationally intensive and might require specialized speech-oriented hardware or careful optimization for real-time applications.
- Cloud vs. Local: Cloud-based solutions (e.g., Google Colab, AWS instances) offer access to powerful hardware without upfront investment but might incur recurring costs. Local hardware allows for full control and can be more economical for very frequent use.

Selecting appropriate generative AI hardware for text, image, and speech applications requires careful consideration. GPUs, particularly high-performance options like the NVIDIA A100, H100, RTX 30, and 40 series, offer versatility for most deep learning tasks. For researchers requiring even greater efficiency or working with extremely large models, specialized AI accelerators like Google TPUs, Graphcore IPUs, or Amazon's Trainium and Inferentia provide tailored solutions. Model size and complexity are key factors in hardware selection. Image generation tasks generally demand higher GPU memory, while generating realistic speech might necessitate specialized hardware or optimization. Additionally, researchers must strategize between cloud-based solutions, offering access to powerful hardware without upfront costs, and investing in local hardware, which grants full control and might be more cost-effective for heavy usage. As the field of AI hardware continues to evolve, staying updated on the latest developments will ensure researchers make the most informed hardware choices for their generative AI projects.

Together, these apps represent a powerful step towards a more inclusive and accessible future. Using the capabilities of AI, developers continue to overcome barriers, making the world more convenient for navigation and understandable to people with visual impairments. As technology continues to evolve, the potential for even more revolutionary visual aid innovations remains on the horizon, promising a future in which accessibility has no bounds.

3. Dynamics of Vision, Text, and Sound in Artificial Intelligence

The pursuit of natural and effective human–computer interactions necessitates an interdisciplinary approach that integrates language, vision, and human behavior. Studies exploring “analysis by synthesis” provide a compelling case study. This methodology leverages Bayesian approaches and predictive models to understand perception, as evidenced by research on deciphering unconventional indirect requests, which rely heavily on pragmatic inferences [28]. Beautemps et al.’s work exemplifies how building detailed models of the human cognitive process, in this case, the role of the temporal lobe in processing indirect requests, can improve the accuracy of the analysis by synthesis approach.

Furthermore, Vinciarelli et al. (2015) highlight the importance of modeling, analyzing, and synthesizing human behavior across diverse interaction scenarios (human–human and human–machine) for designing intuitive AI systems [29]. Their work emphasizes the role of various methodologies, including motion tracking, Computer Vision, and signal processing, in detecting behavioral cues. Additionally, Vinciarelli et al. discuss the need for analytic, descriptive, predictive, and classification models to effectively analyze interpersonal influence. Evaluating these models across standardized datasets and comparing their performance would provide valuable insights into their relative strengths and weaknesses in different contexts.

This emphasis on integrating diverse sensory modalities is further highlighted by Nuske et al.’s (2022) exploration of audio-visual speech synthesis with sensor measurements [30]. Their work exemplifies the need for understanding how humans integrate auditory and visual cues during speech perception. Nuske et al. specifically investigate techniques like the functional modeling of overlapping temporal processes and alternative denoising methods for Mel-frequency cepstral coefficients (MFCCs) and Linear Predictive Cepstral Coefficients (LPCCs)—acoustic features crucial for speech synthesis. Evaluating such methodologies across various datasets and comparing them to established approaches would solidify their effectiveness in generating natural-sounding and engaging speech for human–robot interaction.

In conclusion, this research highlights the promise of interdisciplinary approaches that combine language, vision, and human behavior modeling for advancing human–computer interactions. Further research should prioritize case studies that showcase the effectiveness of specific methodologies and include comparisons across standard datasets to assess their generalisability and performance in various contexts.

In the continuation of the paper, we illustrate the proposed taxonomy of this interdisciplinary field and present concrete approaches in this domain.

We illustrate the overall categorization of reviewed approaches in Figure 2. Due to the high data-type variability involved in this study, we group these methods on the top level on the input–output data-type basis. At the mid-level, we analyze the commonly proposed methodology patterns that are often found in the relevant literature and we organize these terms into colored sub-groups. On the bottom level, we then gather different technique types, where the colored blocks indicate the relation of each technique with the sub-group from the mid-level.

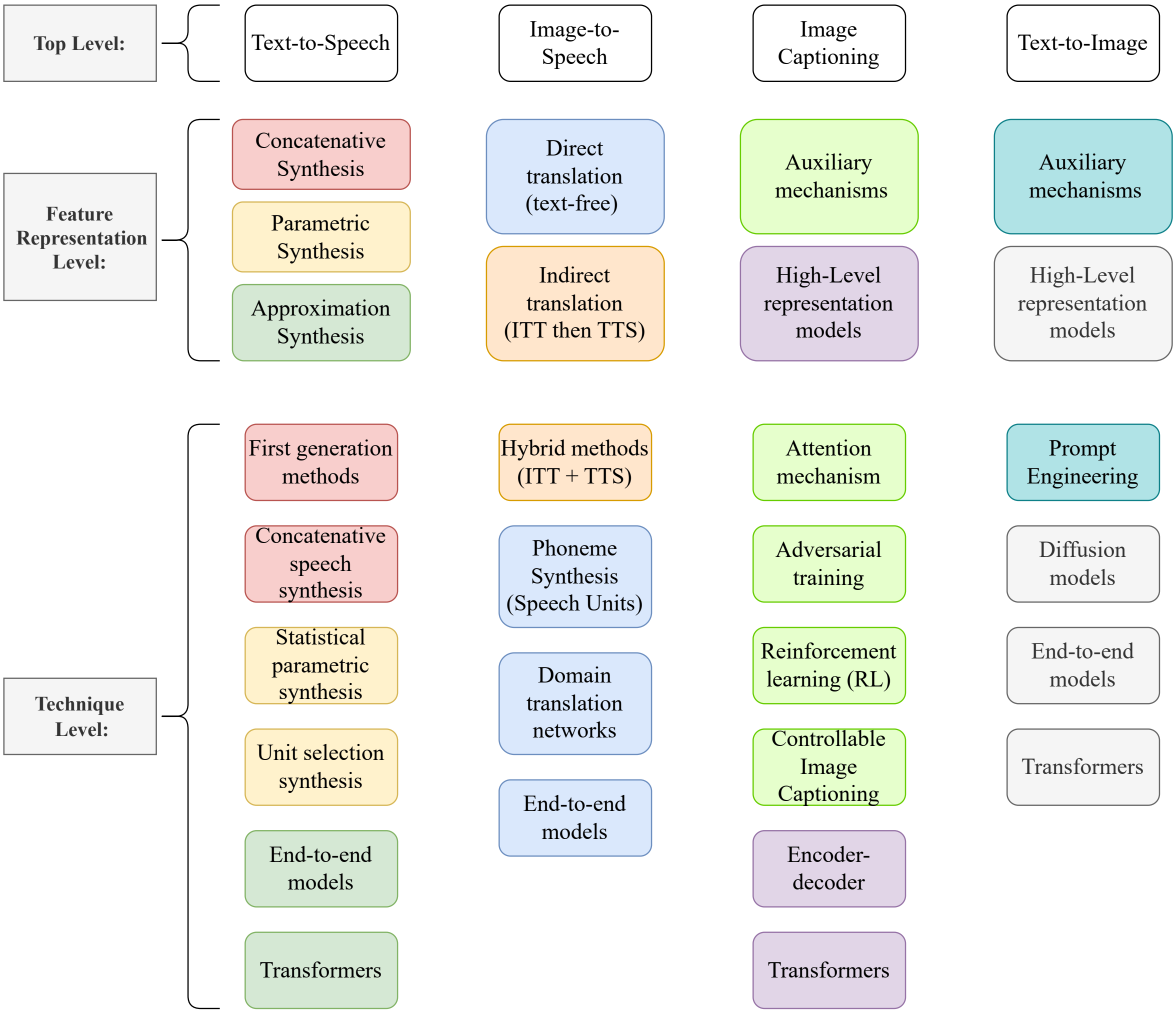

Figure 2. Proposed taxonomy of reviewed techniques operating with text, sound, and vision data. On the top level, we divide the methods based on the produced input–output data pairs, and then in the middle (feature) level, we divide each group into representation-level sub-groups. Finally, we collect and list commonly used techniques in each sub-group, including typically used models, architectures, and mechanisms at the bottom (technique) level. We use the same color annotations in the middle and bottom levels to annotate which techniques at the bottom belong to which mid-level feature representation sub-group.

3.1. Methods and Techniques in Image Captioning

The term “Image-to-Text conversion” encompasses the broader concept of extracting meaning or data from an image. This can involve a range of activities, including object identification, scene understanding, image classification, and the creation of text labels or descriptions appropriate to the content of the image. It also includes all the procedures that convert visual data into textual data.

In the context of this article, image captioning is considered. Image captioning is one of the major research areas in Artificial Intelligence (AI) where image understanding and linguistic description intersect. A fundamental aspect of image understanding involves detecting and recognizing objects, understanding scene type or location, and understanding object properties and their interactions. Forming coherent sentences requires mastery of both the syntactic and semantic aspects of language [31].

Various techniques and methods have been developed to solve the Image Captioning problem. This section will give an overview of methods for creating image captions.

1. The encoder–decoder architecture is the foundation of the majority of image captioning models. To extract visual information from images, Convolutional Neural

Networks (CNNs) are used as encoders, and Recurrent Neural Networks (RNNs) or Transformers are used as decoders to create captions. This framework has been widely used in early image captioning models because of its straightforward and efficient construction [31–35].
Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) are types of RNNs with gating mechanisms to control the flow of information. They address the vanishing gradient problem and are commonly used in sequence-to-sequence tasks such as image captioning, thus effectively capturing sequential dependencies [32].

2. When producing each word in a caption, Attention Mechanisms allow the models to concentrate on particular areas of an image. This enhanced the capacity of the model to represent intricate linkages and features [32,36,37].
3. Transformer-based models have been used for image captioning, motivated by the way transformers perform well in Natural Language Processing (NLP) tasks. They effectively and efficiently capture long-range relationships [34,38,39].
4. Reinforcement Learning was used to fine-tune the image captioning models. The model is trained to maximize the reward signal, which is often computed based on the quality of the generated captions. Moreover, it allows for the optimization of non-differentiable metrics and improves caption quality [32,40–42].
5. Adversarial Training: Generative Adversarial Networks (GANs) are employed to improve the realism and diversity of generated captions. A discriminator was trained to distinguish between real and generated captions [32,42–45].
6. Controllable Image Captioning: Models are designed to generate captions with specific attributes or styles, allowing for control over the content of the generated captions [35,46].

The study of Image-to-Text conversion in AI reveals a complex landscape. Creating image captions, the most important task in this field requires a subtle understanding of image content and linguistic subtleties. Delving deeper into the methodology of caption creation, the underlying encoder–decoder architecture utilizes CNN as an encoder and RNN as a decoder. Attention Mechanisms enhance descriptiveness by focusing on specific regions of the image during word generation. Transformer-based models effectively capture long-range connections, and reinforcement learning optimizes undifferentiated metrics. Adversarial learning using GANs enhances realism, and supervised image captions allow for customized results. This highlights that image captioning is a symphony of techniques harmonizing visual perception and linguistic expression in AI.

3.2. Evolution of Text-to-Speech (TTS) Technologies and Techniques

Several AI techniques are employed for TTS systems, enabling machines to convert written text into spoken words. Text-to-Speech conversion is typically viewed as a two-stage process. First, an abstract basic linguistic representation of the text is created using phonemes, accent symbols, and syntactic structure markers. Subsequently, a speech path model (synthesizer), controlled by a set of rules, converts the sequence of phonemes into sound [47].

In [48], Taylor reviewed the three main synthesis methods, collectively known as first-generation methods, that dominated until the late 1980s. Common problems arise when Formant, Classical Linear Prediction, and Articulatory Synthesis are used. They all rely on the same or comparable sources to produce speech signals with a certain value but do not perform well in recreating natural-sounding speech. The uniqueness of the source, which is influenced by frequency and vocal effort, limits the authenticity of “average” models, making them more difficult to create and acceptable to all speakers. In addition, because of their interdependence, modeling the source and filter as independent entities is insufficient, resulting in complex source models when moving from one speech to another.

Researchers at the now-defunct ATR (Advanced Telecommunications Research) Translational Telephony Laboratory in Kyoto, Japan developed CHATR (Collected Hacks from ATR) in the mid-1990s as a breakthrough technology for Concatenative Speech Synthesis.

This technology was first described in 1994 [49,50], and in 1996 [51,52], a website appeared with a corpus containing 1537 samples of synthesized speech from that era. Although this was not the first Concatenative Speech Synthesis system, it was the first to use raw waveform segments directly, without the use of signal processing. This major change not only simplified the synthesis process but also allowed for the use of very high-quality recordings—some even stereo—that accurately captured all the subtleties of the recorded people's voices and speech patterns. This made it possible to replace the unnatural sound of parametric synthesis and produce a remarkably realistic-sounding speech. Nevertheless, Concatenative TTS requires an impressive collection of recordings to capture all the possible combinations of speech units to produce words. Another disadvantage is that, since combining can decrease the fluency of stress, emotion, prosody, etc., the resulting voice can be less natural and emotional [53].

To overcome the shortcomings of Concatenative TTS, an alternative known as Statistical Parametric Speech Synthesis (SPSS) was developed in the late 1990s [54,55]. Using a stochastic time-series model, SPSS models acoustic features based on data. Known as HMM-based speech synthesis, SPSS uses Hidden Markov Models (HMMs) to encode various linguistic parameters, in addition to phoneme sequences. Similar to unit selection methods, SPSS uses linguistic information to determine acoustic parameters from HMMs to drive a vocoder, which is a rudimentary speech production model that generates speech signals using vocal and excitatory characteristics [56–62]. Compared to previous TTS systems, SPSS offers the following advantages. First, the reproduced sounds are more realistic. Second, it provides versatility in terms of changing the parameters that control generated speech. Third, the number of data required is smaller because fewer recordings are required than in Concatenative Synthesis. However, SPSS has disadvantages; examples of artefacts that can affect the intelligibility of the generated speech are noise or muffled audio. In addition, synthesized speech can sound artificial and merely like an impersonation of human-recorded speech [63].

The first attempts to integrate deep neural networks into SPSS, such as deep neural network (DNN) [63,64] and recurrent neural network (RNN)-based models [65–67], were made in the early 2010s as neural networks and deep learning evolved rapidly [62]. These models, adhering to the SPSS paradigm, use neural networks instead of an HMM to predict acoustic aspects based on linguistic information. Later, Wang et al. [68] proposed the generation of acoustic features directly from phoneme sequences, which was the first study on end-to-end speech synthesis.

Taylor [48] highlighted the prevalence of Unit Selection Synthesis as the leading technique in contemporary TTS systems. Unit selection, an extension of Concatenative Systems, addresses significant challenges in managing large databases of speech units, expanding prosodic control beyond fundamental frequency (F0). It also aims to mitigate distortions that arise from signal processing. Having a diverse set of speech units with variations in prosody and other characteristics for each linguistic type is the fundamental notion behind unit selection. The method is expanded to incorporate variations such as stressed and unstressed versions, phrase-final and non-phrase-final versions, or other linguistic elements like pitch variations, rather than only recording one version for each diphone. The result is a richer database of units. This is a change from Concatenative TTS, which followed the old paradigm of having a single example (unit) for every diphone.

Neural network-based Text-to-Speech synthesis (neural TTS) is a recent innovation in deep learning that builds speech synthesis models based on deep neural networks. The first neural TTS models appeared in 2016 with the release of WaveNet [69], which presented a way to generate speech signals based on direct linguistic features. Thus, speech signals can be generated directly from text using fully end-to-end TTS systems, such as Tacotron [70], ClariNet [71], FastSpeech 2s [72], and EATS [73]. Neural-network-based speech synthesis has significant advantages over earlier TTS systems based on Concatenative and SPSS. These advantages include improved speech quality in terms of naturalness and intelligibility as well as a reduced need for feature development and human speech preprocessing [62].

Transformer models, exemplified by GPT [74,75] and BERT [76], have also made substantial inroads into TTS. Their contextual understanding of language and ability to generate speech based on text inputs mark a promising avenue for more contextually accurate and natural speech synthesis.

AI's capabilities have extended to voice cloning, allowing for the replication and customization of specific voices. By learning from minimal voice samples, AI models can synthesize speech in designated voices, fostering personalization and adaptability [77–79].

These methods have continually evolved, to generate more natural, human-like speech. The combination of deep learning, neural networks, and the application of advanced linguistic and acoustic models has significantly improved the quality of synthesized speech, making TTS systems more efficient and natural-sounding.

3.3. Advancements in Image-to-Speech Systems

Image-to-Speech is a relatively new task that combines image and speech processing. Some of the works on this topic are performed using separate techniques for Image-to-Text and Text-to-Speech [80–83]. However, the task of direct Image-to-Speech conversion is also being worked on.

Hasegawa-Johnson et al. [84] first introduced an Image-to-Speech task to create spoken descriptions of images without relying on text. They divided the process into two stages: the first stage generates speech units (like phonemes) from the image, and the second stage synthesizes speech from these units. Their method heavily relies on image descriptions in terms of sound sequences to train the first stage. They compared three ways of obtaining these sound units, but only the one based on native language phonemes showed reasonable performance, limiting the system's applicability to unwritten languages.

Hsu et al. [80] used an audio-visual model to learn linguistic units from visually grounded speech. They applied these learned speech units to the Image-to-Speech task, achieving reasonable performance. Effendi et al. [85] took a different approach, using a self-supervised model to learn speech unit representations without paired image-speech data. Their model's encoder–decoder architecture allowed it to synthesize speech from predicted speech units by the Image-to-Speech model. Both approaches surpassed the pseudo-phone-based method used in the earlier work.

Xinsheng Wang et al. [86] pioneered an end-to-end method for generating spoken descriptions of images. Their groundbreaking work showcased the feasibility of creating spoken image descriptions without relying on text or intermediate speech units.

In their work on "Audio description from an image by modal translation network", Ning et al. [87] introduced an Image-to-Audio-Description (I2AD) task crucial for various applications. They constructed three substantial audio caption datasets to delve into this task. Their Modal Translation Network (MTNet) aimed to address I2AD by exploiting the inherent relationship between images and audio, translating image information into audio features. This network included an audio generation sub-network utilizing a 1D convolution kernel with holes to model complex phoneme relationships, ensuring natural and understandable audio descriptions.

In a text-free Image-to-Speech process, the approach centers on generating spoken descriptions or interpretations of images without relying on accompanying text. This process typically involves various techniques within Computer Vision and speech synthesis.

The development of such AI systems often involves labeled datasets for training and optimizing the models to perform accurately. These datasets typically contain pairs of images and corresponding text or speech descriptions, allowing the AI to learn the relationship between the visual content and its textual description.

3.4. Image Generation Based on a Text Description

While most of the article was devoted to synthesizing audio descriptions from visual input data, there is growing interest in the inverse problem of generating visual content from natural language descriptions. This area is often called "Text-to-Image generation".

Because an infinite number of images can be associated with a single verbal description, Text-to-Image generation is multimodal. The challenge is to capture different visual interpretations that can arise from the same input text. Deep learning, namely, generative adversarial networks (GANs) and specialized Text-to-Image techniques based on diffusion models, can address this complexity. Transformers play a vital role in this field [88].

GANs are typically used for Text-to-Image generation. In a GAN, the generator creates images using random noise as input, whereas the discriminator assesses how well these created images resemble real images [89].

Agnese et al. summarized GAN-based Text-to-Image synthesis into two types, Simple GAN (Conditional GAN) frameworks and Advanced GAN frameworks, and proposed a taxonomy that divides advanced GAN-based Text-to-Image synthesis frameworks into four categories [90]. The categories are:

- Semantic enhancement GANs (DC-GANs, GAN-INT, GAN-CLS, GAN-INT-CLS, Dong-GAN, Paired-D GAN, and MC-GAN);
- Resolution enhancement GANs (StackGAN, StackGAN++, AttGAN, obj-GANs, HDGAN, and DM-GAN);
- Diversity enhancement GANs (AC-GANs, TAC-GAN, Text-SeGAN, MirrorGAN, and Scene Graph GAN);
- Motion enhancement GANs (ObamaNet, T2S, T2V, and StoryGAN).

Each category addresses a specific aspect of Text-to-Image synthesis and includes common frameworks that use GANs to solve these problems.

GANs have been successful in generating images; however, they suffer from unstable training [91]. Diffusion models offer a significant advantage in Text-to-Image tasks compared with GANs because of their ability to address such issues. Diffusion models offer a more consistent training experience and can produce high-quality digital images with greater variation and quality. Furthermore, diffusion models have been demonstrated to effectively maintain the global content of input images, enabling simultaneous diffusion models capable of learning both explicit information and abstract aesthetics at the same time [92].

Transformer architectures, which were formerly prominent in NLP applications, have spread to the field of Text-to-Image synthesis. Transformers can be used to handle various multimodal tasks, as demonstrated by models such as DALL-E. Transformers excel at capturing contextual nuances by encoding textual information and creating images within a cohesive framework. This allows for a more precise and contextually relevant image synthesis based on textual prompts [93].

As Text-to-Image creation techniques continue to be investigated, one more important component is the use of prompt modifiers. Prompt modifiers play a crucial role in Text-to-Image generation by allowing users to enhance the quality and specificity of generated images. These modifiers enable users to influence various aspects of generated images, including style, content, and overall outcome, by adjusting the text prompt provided to the generative model [94]. The six types of prompt modifiers identified in the taxonomy presented by Oppenlaender [95] are as follows:

- Subject terms: Denotes the subject of the image.
- Style modifiers: Indicates a specific artistic style for the image.
- Image prompts: Provides a reference image to convey the desired style or subject.
- Quality boosters: Terms intended to enhance the quality of generated images.
- Repeating terms: Repetition of subjects or style terms to reinforce desired elements.
- Magic terms: Terms that are semantically different from the rest of the prompt, aiming to produce unexpected or surprising results.

Text-to-Image synthesis is a field that uses a variety of techniques such as transformers, diffusion models, and GANs. While GANs, which are popular but prone to instability during training, have been categorized under several different headings, diffusion models provide steady training and better image quality. Transformers have switched from NLP to Text-to-Image tasks, demonstrating flexibility in models such as DALL-E. Furthermore, prompt modifier integration allows users to fine-tune output images by modifying text prompts and impacting style, content, and quality.

3.5. Ethical Considerations and Potential Unintended Consequences

The technologies presented in the paper, while offering great potential benefits, also raise concerns about unintended consequences. These include the potential for facilitating the spread of deepfakes and misinformation, the perpetuation of biases within the models, threats to privacy and data security, the creation of new accessibility barriers if not designed inclusively, and the possible devaluation of human artistic skills due to the ease of generating realistic content. Proactive measures must be taken to mitigate these risks, including ethical research, bias detection tools, robust privacy frameworks, and inclusive design principles.

3.6. Challenges and Opportunities

The rapid advancement of these AI technologies necessitates the careful consideration of their ethical implications. Further development must prioritize privacy and data security, particularly when handling sensitive personal images and descriptions. Future work will involve the development of robust privacy protocols, including transparent data collection, secure storage, and user-controlled access mechanisms. Additionally techniques like differential privacy and federated learning to protect individual data while training models will most likely be more and more applied in the future. Furthermore, a critical examination of potentially unintended consequences, such as the perpetuation of biases or the ability to synthesize misleading content, will be needed in order to mitigate these risks.

4. Overview of Existing Image Captioning, Text-to-Speech, and Image-to-Speech Datasets

4.1. Image Captioning Datasets

Image captioning datasets are specifically formed to address the task of creating text descriptions or captions for images. This section will give an overview of some image captioning datasets.

The SBU (Stony Brook University) Captioned Photo Dataset is a web-based collection of over 1 million images retrieved from the Internet, each accompanied by visually relevant text descriptions. The dataset was created through a long process involving queries to Flickr with a large number of query term pairs (objects, attributes, actions, things, and scenes). This initial set of photos with corresponding text is then filtered to ensure the relevance and visual expressiveness of the descriptions. Filtering criteria include the observed length of the visual descriptions, the presence of at least two words from predefined term lists, and the inclusion of at least one prepositional word indicating visible spatial relationships (e.g., “at”, “under”). The resulting SBU Captioned Photo Dataset serves as a valuable resource for captioning methods by providing a diverse set of images with visually meaningful textual descriptions [96].

The Flickr8k dataset is a collection of 8092 images from Flickr, each accompanied by five crowdsourced captions. The images in this set are about people or animals performing some action and were selected from six different Flickr groups to depict different scenes and situations. The captions in this dataset are shorter and focus more on the main aspects of the image. The Flickr8k dataset is a unique resource for image description and has been widely used in sentence-based image annotation and search [97].

The Flickr30k dataset is a large-scale benchmark dataset for research in image captioning and multimodal technologies. It consists of 31,783 images collected from the Flickr website, each accompanied by five descriptive captions written by different annotators. The captions are written in English and describe the content of the image from different perspectives, resulting in a diverse set of descriptions for each image. This dataset has been widely used for training and evaluating image captioning models, as well as for research in multimodal learning and Natural Language Processing [98].

The Microsoft Common Objects in COntext (MS COCO) dataset is a large-scale dataset for recognizing, segmenting, and captioning images. It contains 328,000 images with over 2.5 million labeled instances of 91 common object categories. This database was created to achieve excellence in object recognition by addressing the issue of object recognition in the context of the broader issue of scene understanding. Objects are labeled by segmenting each instance to help localize the object accurately. Numerous crowd workers participated in the creation of the dataset and utilized new user interfaces for category detection, instance spotting, and instance segmentation [99].

The Microsoft COCO Captions dataset is a collection of human-created captions for images contained in the MS COCO dataset. The set contains over 1.5 million captions for more than 330,000 images. The captions were collected using Amazon's Mechanical Turk and followed certain guidelines to describe important parts of a scene without including irrelevant details or future/past events [100].

SentiCap dataset is designed to generate sentiment-enriched image descriptions. The crowdsourced dataset involves rewriting image captions based on objective descriptions from MS COCO with the inclusion of affective norms for English words selected by workers. Adjective–noun pairs collected from online image captions are labeled positive or negative sentiments. The dataset is tested for quality using an Amazon Mechanical Turk task that evaluates the descriptiveness and appropriateness of sentiment. By focusing on the viewer's objective emotional response to images, SentiCap provides a unique resource for learning models for creating emotionally expressive image captions, bridging the gap between visual content and semantic connotations in language [101].

The Conceptual Captions dataset is a large-scale database of image caption annotations containing approximately 3.3 million image-caption pairs. Its uniqueness lies in the fact that images and their raw descriptions are collected from the web, providing a wider variety of styles than other curated datasets. The dataset was created programmatically using the Flume pipeline, which processes billions of Internet pages in parallel and includes image and text filtering steps to ensure the resulting captions are clean, informative, fluent, and learnable. This dataset was used to evaluate several image caption creation models [102].

The VizWiz-Captions dataset is the first publicly available dataset dedicated to images taken by blind people. This dataset is based on existing data collected from users of the VizWiz mobile application and consists of 39,181 publicly available images, each accompanied by five captions. Images were sent by users who recorded verbal questions to obtain descriptions of the images or answers to their questions from remote people. Signature crowdsourcing on Amazon Mechanical Turk was used to create the dataset, with an adaptation of a task interface developed in the vision community. The interface encouraged crowdsourcing participants to describe all of the relevant parts of an image to a blind person, avoiding speculation about content and appropriately addressing image quality issues. The dataset addresses real-world image conditions faced by blind photographers and provides a valuable resource for the development of more generalized Computer Vision algorithms. The collection includes redundant captions for quality control, resulting in 195,905 captions post-processed, including spell checking. The goal of the dataset is to contribute to the understanding of real user needs and concerns when creating image captions, beyond the far-fetched parameters of existing datasets [103].

The Localized Narratives dataset is a multimodal image annotation dataset designed to bridge vision and language. In this unique database, annotators describe images using their voice while simultaneously using the mouse to point to areas of interest in the image. The synchronization of voice description and mouse pointing allows each word in the description to be localized to a specific region of the image. This dense visual reference is represented as a segment of the mouse trace for each word, representing a unique and valuable form of annotation. The dataset includes annotations for 849,000 images covering all COCO, Flickr30k, and ADE20K, and 671,000 images from Open Images datasets. The annotations are in the public domain. Each annotation record contains information such as dataset and image identifiers, annotator identifiers, image captions, temporal characteristics of spoken words, mouse movement trace segments, and the relative URL path to the corresponding OGG (Ogging) voice recording. The dataset aims to facilitate research at the interface of vision and language, offering a rich source of information for tasks such as image captioning and the accurate spatial localization of language [104].

The TextCaps dataset is a resource for exploring the task of creating image captions with reading comprehension in mind. The dataset, consisting of 145,000 captions for 28,000 images, aims to address the shortcomings of existing image captioning approaches by focusing on understanding written text and incorporating it into image descriptions. The challenge is to recognize text, relate it to the visual context, and decide how to incorporate text into captions. This involves spatial, semantic, and visual reasoning between several textual tokens and visual entities such as objects in images. The dataset is designed to test the reading abilities of image captioning models and provides an opportunity to train image captioning models to efficiently process and incorporate information from text into images [105].

The LAION (Large-scale Artificial Intelligence Open Network) COCO dataset is the world's largest collection of generated high-quality signatures for publicly available web images, comprising 600 million signatures. Built as an ensemble of BLIP (Bootstrapping Language-Image Pre-training) L/14, CLIP (Contrastive Language–Image Pretraining) (L/14 and RN50x64) versions and a fine-tuned T0 model, the dataset is designed to study the complementarity of synthetic captions with the five billion natural captions in Laion5B. The captions are generated for images from the English-speaking subset of Laion5B and are published in the public domain for research purposes. The method involves several steps, including caption generation, ranking, and grammatical correction. Human experts found that the BLIP and CLIP ensemble can generate signatures with a quality close to human-written MS COCO signatures. The dataset is provided as parquet files including original signatures, URLs, best signatures, and alternative signatures with lower CLIP-like scores. Researchers can download the dataset to study the impact of synthetic captions on trained models and explore the potential of this large-scale captioning resource [106].

Image captioning datasets are crucial for advancing Computer Vision and Natural Language Processing. A comparison of these datasets is shown in Table 1. MS COCO excels in large-scale, segmented images, while Flickr8k and Flickr30k offer insights into sentence-based annotation. Conceptual Captions provides diverse, web-based pairs; VizWiz-Captions focuses on inclusivity; and TextCaps emphasizes reading comprehension. Localized Narratives introduces dense visual grounding, and SentiCap explores sentiment enrichment. SBU Captioned Photo Dataset offers diverse, web-sourced descriptions. LAION COCO, with 600 million synthetic captions, presents a unique opportunity to study complementarity with natural captions. Together, these datasets drive advancements, catering to various vision-language aspects and fostering AI system development.

Table 1. Comparison of image captioning datasets.

Dataset	Source	Volume of Images	Volume of Captions	Annotation Style	Purpose (Usage)
SBU Captioned Photo Dataset [96]	Web-based, Flickr	Over 1 million	Over 1 million	Visual relevance, filtering	General image captioning
Flickr8k [97]	Flickr	8092	8092 × 5	Crowdsourced, main aspects	Sentence-based image annotation, search
Flickr30k [98]	Flickr	31,783	31,783 × 5	English captions, diverse	Image captioning, multimodal learning, and natural language processing

Table 1. *Cont.*

Dataset	Source	Volume of Images	Volume of Captions	Annotation Style	Purpose (Usage)
MS COCO [99]	Internet, crowd workers	328,000	Over 2.5 million labeled instances	Object recognition, and segmentation	Image captioning, object recognition, and segmentation
MS COCO Captions [100]	MS COCO dataset	over 330,000	Over 1.5 million	Human-created, guidelines	Caption quality, scene description
SentiCap [101]	MS COCO dataset	Several thousand	Over 2000	Sentiment-enriched captions	Emotionally expressive image captions
Conceptual Captions [102]	Web-based, Flume pipeline	Approximately 3.3 million	approximately 3.3 million	Image and text filtering, diverse styles	Evaluating Image caption creation models
VizWiz-Captions [103]	VizWiz mobile app	39,181	39,181 × 5	Captions for blind users	Real-world image conditions for blind photographers
Localized Narratives [104]	MS COCO, Flickr30k, ADE20K, Open Images	849,000 + 671,000		Voice descriptions with mouse traces	Vision and language research, image captioning
TextCaps [105]	Open Images v3 dataset	28,000	145,000	OCR system and human annotators	Reading abilities of image captioning models
LAION COCO [106]	Publicly available web-images, English subset of Laion-5B	600 million	600 million	Synthetic captions	Large-scale captioning resource, complementarity study

4.2. Text-to-Speech Datasets

Text-to-speech (TTS) systems rely heavily on diverse datasets for training and evaluating the performance of voice synthesis models. Several prominent datasets have emerged as crucial resources in this domain, each offering unique characteristics and diverse speech samples.

The LJ Speech Dataset is a public domain collection for TTS research. It includes 13,100 short audio clips of a single speaker reading excerpts from non-fiction books published between 1884 and 1964. The dataset totals about 24 h and includes transcriptions and normalized transcriptions that expand numbers and units to full words. The audio, recorded in 2016–2017 as part of the LibriVox project, is presented in single-channel 16-bit PCM (Pulse Code Modulation) WAV format with a sampling rate of 22,050 Hz. Each 1 to 10-s clip is segmented by silence in the recording. In the dataset: total number of words (225,715), total number of characters (1,308,678), number of individual words (13,821). Notably, some abbreviations in the text are expanded, and 19 transcriptions contain non-ASCII characters [107].

The LibriTTS corpus, designated as SLR60, is a significant resource in the field of speech research, licensed under CC BY 4.0. Designed for TTS research, it includes approximately 585 h of English speech sampled at 24 kHz. Created by Heiga Zen in collaboration with the Google Speech and Google Brain teams, LibriTTS is an extension of the LibriSpeech corpus [108] derived from mp3 audio files (LibriVox) and text files (Project Gutenberg). Features include a 24 kHz sampling rate, sentence-level segmentation, the inclusion of both original and normalized texts, and the exclusion of utterances with significant background noise. In addition, LibriTTS features the ability to extract contextual information from neighboring sentences, which enhances its suitability for prospective TTS research [109].

The RyanSpeech corpus is a speech dataset designed for research on automated TTS systems, fulfilling the need for a high-quality publicly available corpus of male speech in a conversational setting. The set contains over 10 h of speech recorded at 44.1 kHz by a professional male voice actor, making it suitable for developing TTS systems in real-world applications. RyanSpeech is the first publicly available TTS corpus with a male voice actor in a conversational environment. Its creation was carefully crafted, involving data collection from various text resources; sentence segmentation; text normalization; and post-processing steps such as silence pruning and audio amplitude normalization. The corpus includes sentences from the Ryan Chatbot, Taskmaster-2, and LibriTTS datasets, covering a variety of conversation topics. The dataset is provided under a CC BY-NC-ND (Creative Commons Attribution NonCommercial NoDerivs) license along with trained models to facilitate further research and development on TTS [110].

The SOMOS (Samsung Open Mean Opinion Scores) dataset presented by Giorgia Magnati et al. is a pioneering contribution to the evaluation of neural TTS synthesis. It is the first extensive Mean Opinion Scores (MOS) dataset specifically designed for the evaluation of state-of-the-art TTS systems. Composed of 20,000 synthetic utterances derived from the widely used LJ Speech dataset, SOMOS includes a diverse set of TTS systems (from 001 to 200) with different acoustic models and prosody. The dataset utilizes the LPCNet vocoder to evaluate the naturalness of speech to ensure the consistency of acoustic features. Example sentences illustrating problems in aspects such as prosody, rhythm, accent, pauses, and pronunciation provide valuable information. Crowdsourced MOS evaluations on Amazon Mechanical Turk contribute to the reliability of the dataset. SOMOS serves as a critical resource to stimulate progress in TTS synthesis evaluation and refine the evaluation of acoustic models [111].

The CVSS (Common Voice Speech-to-Speech) dataset is a feature-rich and comprehensive corpus of multilingual-to-English Speech-to-Speech translations. CVSS, which includes sentence-level parallel translations from 21 languages into English, is based on the Common Voice speech corpus and the CoVoST 2 Speech-to-Text translation dataset. Two versions are available for each language pair: CVSS-C contains synthetic translations performed by a sequential canonical speaker's voice, which provides highly natural and clean translations and is ideal for user-oriented applications. In contrast, CVSS-T provides translated speeches with voices transferred from the source speeches, which preserves the similarity of voices across languages. The dataset of about 1900 h of speech is synthesized using state-of-the-art Text-to-Speech models trained on the LibriTTS corpus. Along with the translated speech, the CVSS includes normalized translation text, which helps in training the models and standardizing the scores. This corpus serves as a valuable resource for advancing research in multilingual Speech-to-Speech translation, offering different approaches to modeling and preserving the speaker's voice in different languages [112].

The AISHELL-3 corpus is a large-scale and highly accurate multilingual Mandarin speech dataset designed for training multilingual TTS systems. The corpus contains about 85 h of emotionally neutral recordings from 218 Mandarin speakers totalling 88, 035 utterances and includes explicit annotation of auxiliary attributes such as gender, age group, and native accent. With professional speech annotation and careful transcription quality control, the transcription accuracy exceeds 98%. The corpus serves as a valuable resource for developing robust synthesis models, and the underlying system includes a speaker verification model to achieve high voice similarity [113].

HUI-Audio-Corpus-German is a significant open-source dataset designed for TTS applications and aims to address the inherent weaknesses of existing datasets. The dataset is created using a processing pipeline that emphasizes high-quality audio and transcription alignment to reduce the manual labor required to create the dataset. Achievements include meeting a minimum duration of 20 h per speaker, providing a sampling rate of 44.1 kHz, normalizing text and audio loudness, and keeping the average audio length within specified limits. In addition, the dataset includes punctuation related to pronunciation and preserves capitalization in transcripts. Special attention to text normalization is paid via the automatic checking of digits, abbreviations, and special characters, as well as the careful manual analysis of transcript samples. The characteristics of the dataset make it a valuable resource for the development of TTS research, especially for the German language, under strict quality and harmonization requirements [114].

The KazakhTTS dataset is a high-quality open-source speech synthesis dataset for the Kazakh language. Developed to address the problem of language resource scarcity, the dataset contains about 42,000 segments totalling 93 h of transcribed audio recordings made by two professional speakers (a female and a male). This is the first publicly available large-scale dataset designed to advance Kazakh TTS applications in both academic and industrial settings. To create the dataset, texts were collected, read, and carefully segmented with audio and text alignment by native Kazakh speakers using the Praat toolkit. The speakers, selected by listening, recorded the articles in a relaxed environment at their natural pace, following orthoepic rules. Evaluation with end-to-end TTS models demonstrated the reliability of the dataset, providing an MOS above 4 for both speakers. The dataset, training recipe, and pre-trained TTS models are freely available [115].

There are also well-known datasets like LibriSpeech [108] and Mozilla Common Voice [116], which are known primarily as Automatic Speech Recognition (ASR) datasets rather than TTS datasets. Despite this, in some cases, these datasets can be used for TTS research.

The discussed Text-to-Speech datasets play a key role in the development of speech synthesis research and development. A comparison of these datasets is shown in Table 2. Each dataset, whether for specific languages, communication scenarios, or evaluation purposes, contributes to the collective development of TTS technology, offering valuable resources for both researchers and developers.

Table 2. Comparison of Text-to-Speech datasets.

Dataset	Source	Volume	Purpose (Usage)
LJ Speech [107]	Non-fiction books published between 1884 and 1964, LibriVox project	24 h	Research in TTS, voice synthesis models
LibriTTS (SLR60) [109]	LibriSpeech, Project Gutenberg	585 h	TTS research, contextual information extraction
RyanSpeech [110]	Ryan Chatbot, Taskmaster-2, and LibriTTS datasets	10 h	Development of TTS systems
SOMOS [111]	Derived from LJ Speech, LPCNet vocoder	20,000 synthetic utterances	Evaluation of TTS synthesis, refinement of models
CVSS (Common Voice Speech-to-Speech) [112]	Common Voice speech corpus, CoVoST 2 Speech-to-Text translation dataset	1900 h	Multilingual Speech-to-Speech translation
AISHELL-3 [113]	Emotionally neutral recordings from 218 Mandarin speakers	85 h	Training multilingual TTS systems
HUI-Audio-Corpus-German [114]	LibriVox	Minimum 20 h per speaker	TTS research, especially for German
KazakhTTS [115]	Manually extracted articles from news websites	93 h	Advancing Kazakh TTS applications

4.3. Image-to-Speech Datasets

As mentioned in the section on Image-to-Speech techniques, most of the works are carried out using a combination of Image-to-Text and Text-to-Speech techniques, and the same is true for datasets. Nevertheless, some datasets can be used for direct Image-to-Speech tasks.

Places Audio Captions: Featuring images with associated audio captions describing scenes and objects [117].

Spoken ObjectNet: Contains images with corresponding spoken descriptions focusing on object recognition and scene understanding [118].

Localized Narratives: Provides images with localized, descriptive narrations capturing various aspects of scenes [104].

Flickr Audio Caption: Annotated images from Flickr paired with human-recorded audio descriptions [119].

SpokenCOCO: Images from the COCO dataset with spoken captions describing the visual content [80].

These datasets serve as fundamental resources, each emphasizing a specific aspect of the Image-to-Speech paradigm. A comparison of Image-to-Speech datasets is shown in Table 3. They enable the training and evaluation of models aimed at converting visual data into comprehensive auditory descriptions, catering to diverse scenes, objects, and contextual information.

Table 3. Comparison of Image-to-Speech datasets.

Dataset	Source	Volume of Spoken Captions	Purpose (Usage)
Places Audio Captions [117]	Places 205 image dataset	Over 400k	Image-to-Speech tasks, scene understanding
Spoken ObjectNet [118]	ObjectNet dataset	50,273	Image-to-Speech tasks, object recognition
Localized Narratives [104]	MS COCO, Flickr30k, ADE20K, Open Images	849,000 + 671,000	Research at the interface of vision and language
Flickr Audio Caption [119]	Flickr 8k	40,000	Image-to-Speech tasks, diverse image descriptions
SpokenCOCO [80]	MS COCO	Approximately 600,000	Image-to-Speech tasks, diverse image descriptions

5. Conclusions

To summarize, the article examines the complex process of converting images to audio descriptions in the fields of image captioning, Text-to-Speech synthesis, and Image-to-Speech conversion. While the focus was primarily on generating audio descriptions from visual data, a brief review of the inverse process, known as “Text-to-Image generation”, was also conducted. The key findings highlight the variety of methodologies used, ranging from traditional encoder–decoder architectures to advanced techniques involving transformers and adversarial learning. The study of datasets such as MS COCO, LibriTTS, and Localized Narratives highlights their key role in training and evaluating models for image captioning and Text-to-Speech synthesis. In addition, the discussion of the tasks of directly converting images into speech and generating images from textual descriptions reveals promising developments in this rapidly growing field.

The value of the conversion coefficient of images into audio descriptions for the future is huge. The integration of advanced technologies, including transformer models and neural networks, suggests a move towards more natural, contextually rich sound descriptions. The emergence of direct Image-to-Speech tasks opens up new opportunities for creating comprehensive and intuitive auditory representations of visual content. This evolution has significant implications for accessibility, improving the user experience for people with visual impairments, and expanding the scope of Artificial Intelligence (AI) applications in various fields.

To move this field forward, researchers are encouraged to explore new architectures that take advantage of transformer models and neural networks for more accurate and contextually rich transformation of images into audio descriptions. As the tasks of direct Image-to-Speech conversion gain momentum, there is a need for data sets specially designed for this purpose, reflecting the nuances of translating visual information into natural, expressive auditory descriptions.

In addition, interdisciplinary collaboration between Computer Vision and Natural Language Processing communities can contribute to a holistic understanding of the challenges and opportunities of converting images into audio descriptions. Ethical considerations, especially in datasets and models, should be the focus to ensure the responsible development and implementation of AI systems.

In conclusion, the transition from images to audio descriptions is constantly evolving, promising a future in which AI systems seamlessly transform visual content into meaningful and inclusive auditory experiences. By solving current problems and exploring innovative approaches, this area can make a significant contribution to both technological progress and social well-being.

Author Contributions: Conceptualization, L.O., Ž.E., and B.M.; methodology, L.O.; validation, B.M., Ž.E.; formal analysis, L.O.; investigation, L.O.; resources, L.O.; data curation, L.O.; writing—original draft preparation, L.O.; writing—review and editing, B.M. and Ž.E.; validation, B.M. and Ž.E.; visualization, B.M. and Ž.E.; supervision, P.P., B.R., B.M. and Ž.E.; project administration, P.P. and B.R.; and funding acquisition, P.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in parts by the ARIS Research Programme P2-0214 “Computer Vision”.

Data Availability Statement: Data sharing is not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. World Health Organization. Blindness and Vision Impairment. 2023. Available online: https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment (accessed on 13 October 2023).
2. Sri, K.S.; Mounika, C.; Yamini, K. Audiobooks that converts Text, Image, PDF-Audio & Speech-Text: For physically challenged & improving fluency. In Proceedings of the 2022 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 20–22 July 2022; pp. 83–88.
3. Unlocking Communication: The Power of Audio Description in Overcoming Language Barriers | Acadestudio. Available online: https://www.acadestudio.com/blog/how-audio-description-is-breaking-down-language-barriers/ (accessed on 21 January 2024).
4. Pashler, H.; McDaniel, M.; Rohrer, D.; Bjork, R. Learning styles: Concepts and evidence. *Psychol. Sci. Public Interest* **2008**, *9*, 105–119. [CrossRef] [PubMed]
5. Moens, M.F.; Pastra, K.; Saenko, K.; Tuytelaars, T. Vision and language integration meets multimedia fusion. *IEEE Multimed.* **2018**, *25*, 7–10. [CrossRef]
6. Guo, J.; He, H.; He, T.; Lausen, L.; Li, M.; Lin, H.; Shi, X.; Wang, C.; Xie, J.; Zha, S.; et al. Gluoncv and gluonnlp: Deep learning in Computer Vision and natural language processing. *J. Mach. Learn. Res.* **2020**, *21*, 845–851.
7. Mogadala, A.; Kalimuthu, M.; Klakow, D. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *J. Artif. Intell. Res.* **2021**, *71*, 1183–1317. [CrossRef]
8. Kleege, G. 7 Audio Description Described. In *More than Meets the Eye: What Blindness Brings to Art*; Oxford University Press: Oxford , UK, 2018. [CrossRef]
9. Snyder, J. *The Visual Made Verbal: A Comprehensive Training Manual and Guide to the History and Applications of audio Description*; Academic Publishing: San Diego, CA, USA, 2020.
10. Snyder, J. Audio description guidelines and best practices. In *American Council of the Blind's Audio Description Project*; American Council of the Blind: Alexandria, VA, USA, 2010.
11. Bittner, H. Audio description guidelines: A comparison. *New Perspect. Transl.* **2012**, *20*, 41–61.
12. Massiceti, D. Computer Vision and Natural Language Processing for People with Vision Impairment. Ph.D. Thesis, University of Oxford, Oxford, UK, 2019.
13. Microsoft Corporation. Seeing AI. Available online: https://www.microsoft.com/en-us/ai/seeing-ai (accessed on 3 November 2023).
14. Envision. Envision—Perceive Possibility. Available online: https://www.letsenvision.com/ (accessed on 3 November 2023).
15. CloudSight, Inc.. TapTapSee—Blind and Visually Impaired Assistive Technology—Powered by CloudSight.ai Image Recognition API. Available online: https://www.taptapseeapp.com (accessed on 3 November 2023).
16. GAATES, the Global Alliance for Accessible Technologies and Environments. Aipoly App Opens Up the World for People with Vision Disabilities. 2017. Available online: https://globalaccessibilitynews.com/2017/03/28/aipoly-app-opens-up-the-world-for-people-with-vision-disabilities/ (accessed on 3 November 2023).

17. Turkel, A. iDentifi. Available online: https://www.getidentifi.com (accessed on 3 November 2023).
18. BlindSquare. Available online: https://www.blindsquare.com/ (accessed on 3 November 2023).
19. We're Aira, a Visual Interpreting Service. Available online: https://aira.io/ (accessed on 3 May 2023).
20. NoorCam. NoorCam MyEye. Available online: https://www.noorcam.com/en-ae/noorcam-myeye (accessed on 3 November 2023).
21. Be My Eyes. Be My Eyes—See the world together. Available online: https://www.bemyeyes.com/ (accessed on 3 November 2023).
22. Lookout—Assisted Vision—Apps on Google Play. Available online: https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.reveal&hl=en%5Ctextunderscore%7B%7DUS&pli=1 (accessed on 3 November 2023).
23. Cyber Timez, Inc.. Cyber Timez. Available online: https://www.cybertimez.com (accessed on 3 November 2023).
24. Eyesynth—Visión a través del oído. Available online: https://eyesynth.com/ (accessed on 4 April 2024).
25. eSight—Electronic Eyewear for the Visually Impaired. 2023. Available online: https://www.esighteyewear.com (accessed on 3 April 2024).
26. GiveVision. Available online: https://www.givevision.net (accessed on 3 April 2024).
27. NuEyes—Empowering Your Vision. Available online: https://www.nueyes.com/ (accessed on 4 April 2024).
28. Beautemps, D.; Schwartz, J.L.; Sato, M. Analysis by Synthesis: A (Re-)Emerging Program of Research for Language and Vision. *Biolinguistics* **2010**, *4*, 287–300. [CrossRef]
29. Vinciarelli, A.; Pantic, M.; Bourlard, H. Open Challenges in Modelling, Analysis and Synthesis of Human Behaviour in Human–Human and Human–Machine Interactions. *Cogn. Comput.* **2015**, *7*, 397–413. [CrossRef]
30. Ashok, K.; Ashraf, M.; Thimmia Raja, J.; Hussain, M. Z.; Singh, D. K.; Haldorai, A. Collaborative analysis of audio-visual speech synthesis with sensor measurements for regulating human–robot interaction. *Int. J. Syst. Assur. Eng. Manag.* **2022**, 1–8. [CrossRef]
31. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 652–663. [CrossRef] [PubMed]
32. Hossain, M.Z. Deep Learning Techniques for Image Captioning. Ph.D. Thesis, Murdoch University, Perth, WA, Australia, 2020.
33. Seshadri, M.; Srikanth, M.; Belov, M. Image to language understanding: Captioning approach. *arXiv* **2020**, arXiv:2002.09536.
34. Chen, F.; Li, X.; Tang, J.; Li, S.; Wang, T. A Survey on Recent Advances in Image Captioning. *J. Phys. Conf. Ser.* **2021**, *1914*, 012053. [CrossRef]
35. Wang, C.; Zhou, Z.; Xu, L. An integrative review of image captioning research. *J. Phys. Conf. Ser.* **2021**, *1748*, 042060. [CrossRef]
36. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 7–9 July 2015; pp. 2048–2057.
37. Jin, J.; Fu, K.; Cui, R.; Sha, F.; Zhang, C. Aligning where to see and what to tell: Image caption with region-based attention and scene factorization. *arXiv* **2015**, arXiv:1506.06272.
38. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 5998–6008.
39. Li, G.; Zhu, L.; Liu, P.; Yang, Y. Entangled transformer for image captioning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8928–8937.
40. Ren, Z.; Wang, X.; Zhang, N.; Lv, X.; Li, L.J. Deep reinforcement learning-based image captioning with embedding reward. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 290–298.
41. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-critical sequence training for image captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7008–7024.
42. Yan, S.; Wu, F.; Smith, J.S.; Lu, W.; Zhang, B. Image captioning using adversarial networks and reinforcement learning. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 248–253.
43. Dai, B.; Fidler, S.; Urtasun, R.; Lin, D. Towards diverse and natural image descriptions via a conditional gan. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2970–2979.
44. Shetty, R.; Rohrbach, M.; Anne Hendricks, L.; Fritz, M.; Schiele, B. Speaking the same language: Matching machine to human captions by adversarial training. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4135–4144.
45. Amirian, S.; Rasheed, K.; Taha, T.R.; Arabnia, H.R. Image captioning with generative adversarial network. In Proceedings of the 2019 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 5–7 December 2019; pp. 272–275.
46. Cornia, M.; Baraldi, L.; Cucchiara, R. Show, control and tell: A framework for generating controllable and grounded captions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8307–8316.
47. Klatt, D. The Klattalk text-to-speech conversion system. In Proceedings of the ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing, Paris, France, 3–5 May 1982; Volume 7, pp. 1589–1592.
48. Taylor, P. *Text-to-Speech Synthesis*; Cambridge University Press: Cambridge, UK, 2009.

49. Black, A.W.; Taylor, P.A. CHATR: A generic speech synthesis system. In Proceedings of COLING-94, Kyoto, Japan, 5–9 August 1994; Volume 2, pp. 983–986.
50. Campbell, N. Prosody and the selection of units for concatenative synthesis. In Proceedings of the ESCA/IEEE 2nd Workshop on Speech Synthesis, New Paltz, NY, USA, 12–15 September 1994; pp. 61–64.
51. Hunt, A.J.; Black, A.W. Unit selection in a concatenative speech synthesis system using a large speech database. In Proceedings of the 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, Atlanta, GA, USA, 9 May 1996; Volume 1, pp. 373–376.
52. Campbell, N. CHATR: A high-definition speech re-sequencing system. In Proceedings of the 3rd Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan, Honolulu, HI, USA, 2–6 December 1996.
53. Tan, X.; Qin, T.; Soong, F.; Liu, T.Y. A survey on neural speech synthesis. *arXiv* **2021**, arXiv:2106.15561.
54. Yoshimura, T.; Tokuda, K.; Masuko, T.; Kobayashi, T.; Kitamura, T. Duration modeling for HMM-based speech synthesis. In Proceedings of the ICSLP, Sydney, NSW, Australia, 30 November–4 December 1998; Volume 98, pp. 29–32.
55. Yoshimura, T.; Tokuda, K.; Masuko, T.; Kobayashi, T.; Kitamura, T. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In Proceedings of the Sixth European Conference on Speech Communication and Technology, Budapest, Hungary, 5–9 September 1999.
56. Tokuda, K.; Yoshimura, T.; Masuko, T.; Kobayashi, T.; Kitamura, T. Speech parameter generation algorithms for HMM-based speech synthesis. In Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 00CH37100), Istanbul, Turkey, 5–9 June 2000; Volume 3, pp. 1315–1318.
57. Yoshimura, T.; Tokuda, K.; Masuko, T.; Kobayashi, T.; Kitamura, T. Mixed excitation for HMM-based speech synthesis. In Proceedings of the Seventh European conference on speech Communication and Technology, Aalborg, Denmark, 3–7 September 2001.
58. Zen, H.; Toda, T.; Tokuda, K. The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006. *IEICE Trans. Inf. Syst.* **2008**, *91*, 1764–1773. [CrossRef]
59. Zen, H.; Tokuda, K.; Masuko, T.; Kobayashi, T.; Kitamura, T. Hidden semi-Markov model based speech synthesis. In Proceedings of the Eighth International Conference on Spoken Language Processing, Jeju Island, Republic of Korea, 4–8 October 2004.
60. Tokuda, K.; Zen, H.; Black, A.W. An HMM-based speech synthesis system applied to English. In Proceedings of the IEEE Speech Synthesis Workshop, Santa Monica, CA, USA, 13 September 2002; pp. 227–230.
61. Black, A.W.; Zen, H.; Tokuda, K. Statistical parametric speech synthesis. In Proceedings of the 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07, Honolulu, HI, USA, 15–20 April 2007; Volume 4, pp. 1229–1232.
62. Zen, H.; Tokuda, K.; Black, A.W. Statistical parametric speech synthesis. *Speech Commun.* **2009**, *51*, 1039–1064. [CrossRef]
63. Zen, H.; Senior, A.; Schuster, M. Statistical parametric speech synthesis using deep neural networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–31 May 2013; pp. 7962–7966.
64. Qian, Y.; Fan, Y.; Hu, W.; Soong, F.K. On the training aspects of deep neural network (DNN) for parametric TTS synthesis. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 3829–3833.
65. Fan, Y.; Qian, Y.; Xie, F.L.; Soong, F.K. TTS synthesis with bidirectional LSTM based recurrent neural networks. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014.
66. Zen, H. Acoustic modeling in statistical parametric speech synthesis-from HMM to LSTM-RNN. In Proceedings of the The First, International Workshop on Machine Learning in Spoken Language Processing (MLSLP2015), Aizu, Japan, 19–20 September 2015; pp. 125–132.
67. Zen, H.; Sak, H. Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 4470–4474.
68. Wang, W.; Xu, S.; Xu, B. First Step Towards End-to-End Parametric TTS Synthesis: Generating Spectral Parameters with Neural Attention. In Proceedings of the Interspeech, San Francisco, CA, USA, 8–12 September 2016; pp. 2243–2247.
69. Oord, A.v.d.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv* **2016**, arXiv:1609.03499.
70. Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Towards end-to-end speech synthesis. *arXiv* **2017**, arXiv:1703.10135.
71. Ping, W.; Peng, K.; Chen, J. Clarinet: Parallel wave generation in end-to-end text-to-speech. *arXiv* **2018**, arXiv:1807.07281.
72. Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; Liu, T.Y. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv* **2020**, arXiv:2006.04558.
73. Donahue, J.; Dieleman, S.; Bińkowski, M.; Elsen, E.; Simonyan, K. End-to-end adversarial text-to-speech. *arXiv* **2020**, arXiv:2006.03575.
74. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
75. Stephenson, B.; Hueber, T.; Girin, L.; Besacier, L. Alternate Endings: Improving prosody for incremental neural tts with predicted future text input. *arXiv* **2021**, arXiv:2102.09914.

76. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
77. Luong, H.T.; Yamagishi, J. Nautilus: A versatile voice cloning system. *IEEE/Acm Trans. Audio Speech Lang. Process.* **2020**, *28*, 2967–2981. [CrossRef]
78. Ruggiero, G.; Zovato, E.; Di Caro, L.; Pollet, V. Voice cloning: A multi-speaker text-to-speech synthesis approach based on transfer learning. *arXiv* **2021**, arXiv:2102.05630.
79. Arik, S.; Chen, J.; Peng, K.; Ping, W.; Zhou, Y. Neural voice cloning with a few samples. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 10019–10029.
80. Hsu, W.N.; Harwath, D.; Song, C.; Glass, J. Text-free image-to-speech synthesis using learned segmental units. *arXiv* **2020**, arXiv:2012.15454.
81. Stephen, O.; Mishra, D.; Sain, M. Real time object detection and multilingual speech synthesis. In Proceedings of the 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), Kanpur, India, 6–8 July 2019; pp. 1–3.
82. Ma, S.; McDuff, D.; Song, Y. Unpaired image-to-speech synthesis with multimodal information bottleneck. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 7598–7607.
83. Bourbakis, N. Automatic Image-to-Text-to-Voice Conversion for Interactively Locating Objects in Home Environments. In Proceedings of the 2008 20th IEEE International Conference on Tools with Artificial Intelligence, Dayton, OH, USA, 3–5 November 2008; Volume 2, pp. 49–55.
84. Hasegawa-Johnson, M.; Black, A.; Ondel, L.; Scharenborg, O.; Ciannella, F. Image2speech: Automatically generating audio descriptions of images. *Casablanca* **2017**, *2017*, 65.
85. Effendi, J.; Sakti, S.; Nakamura, S. End-to-end image-to-speech generation for untranscribed unknown languages. *IEEE Access* **2021**, *9*, 55144–55154. [CrossRef]
86. Wang, X.; Van Der Hout, J.; Zhu, J.; Hasegawa-Johnson, M.; Scharenborg, O. Synthesizing spoken descriptions of images. *IEEE/Acm Trans. Audio Speech Lang. Process.* **2021**, *29*, 3242–3254. [CrossRef]
87. Ning, H.; Zheng, X.; Yuan, Y.; Lu, X. Audio description from image by modal translation network. *Neurocomputing* **2021**, *423*, 124–134. [CrossRef]
88. Ivezić, D.; Bagić Babac, M. Trends and Challenges of Text-to-Image Generation: Sustainability Perspective. *Croat. Reg. Dev. J.* **2023**, *4*, 56–78. [CrossRef]
89. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680.
90. Agnese, J.; Herrera, J.; Tao, H.; Zhu, X. A survey and taxonomy of adversarial neural networks for text-to-image synthesis. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2020**, *10*, e1345. [CrossRef]
91. Jabbar, A.; Li, X.; Omar, B. A survey on generative adversarial networks: Variants, applications, and training. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–49. [CrossRef]
92. Zhang, C.; Zhang, C.; Zhang, M.; Kweon, I.S. Text-to-image diffusion model in generative ai: A survey. *arXiv* **2023**, arXiv:2303.07909.
93. DALL·E: Creating Images from Text. Available online: https://openai.com/research/dall-e (accessed on 16 February 2024).
94. Liu, V.; Chilton, L.B. Design guidelines for prompt engineering text-to-image generative models. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 30 April–6 May 2022; pp. 1–23.
95. Oppenlaender, J. A taxonomy of prompt modifiers for text-to-image generation. *Behav. Inf. Technol.* **2023**, 1–14. [CrossRef]
96. Ordonez, V.; Kulkarni, G.; Berg, T. Im2text: Describing images using 1 million captioned photographs. *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 1143–1151.
97. Hodosh, M.; Young, P.; Hockenmaier, J. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res.* **2013**, *47*, 853–899. [CrossRef]
98. Young, P.; Lai, A.; Hodosh, M.; Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.* **2014**, *2*, 67–78. [CrossRef]
99. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Part V 13, pp. 740–755.
100. Chen, X.; Zitnick, C.L. Learning a recurrent visual representation for image caption generation. *arXiv* **2014**, arXiv:1411.5654.
101. Mathews, A.; Xie, L.; He, X. Senticap: Generating image descriptions with sentiments. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; Volume 30.
102. Sharma, P.; Ding, N.; Goodman, S.; Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, VIC, Australia, 15–20 July 2018; pp. 2556–2565.
103. Gurari, D.; Zhao, Y.; Zhang, M.; Bhattacharya, N. Captioning images taken by people who are blind. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Part XVII 16, pp. 417–434.

104. Pont-Tuset, J.; Uijlings, J.; Changpinyo, S.; Soricut, R.; Ferrari, V. Connecting vision and language with localized narratives. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Part V 16, pp. 647–664.
105. Sidorov, O.; Hu, R.; Rohrbach, M.; Singh, A. Textcaps: A dataset for image captioning with reading comprehension. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Part II 16, pp. 742–758.
106. Schuhmann, C.; Köpf, A.; Vencu, R.; Coombes, T.; Beaumont, R. Laion Coco: 600M Synthetic Captions From Laion2B-en | LAION. 2022. Available online: https://laion.ai/blog/laion-coco/ (accessed on 5 November 2023).
107. Ito, K.; Johnson, L. The lj speech dataset 2017. Available online: https://keithito.com/LJ-Speech-Dataset/ (accessed on 5 November 2023).
108. Panayotov, V.; Chen, G.; Povey, D.; Khudanpur, S. Librispeech: An asr corpus based on public domain audio books. In Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), South Brisbane, QLD, Australia, 19–24 April 2015; pp. 5206–5210.
109. Zen, H.; Dang, V.; Clark, R.; Zhang, Y.; Weiss, R.J.; Jia, Y.; Chen, Z.; Wu, Y. Libritts: A corpus derived from librispeech for text-to-speech. *arXiv* **2019**, arXiv:1904.02882.
110. Zandie, R.; Mahoor, M.H.; Madsen, J.; Emamian, E.S. Ryanspeech: A corpus for conversational text-to-speech synthesis. *arXiv* **2021**, arXiv:2106.08468.
111. Maniati, G.; Vioni, A.; Ellinas, N.; Nikitaras, K.; Klapsas, K.; Sung, J.S.; Jho, G.; Chalamandaris, A.; Tsiakoulis, P. SOMOS: The Samsung Open MOS Dataset for the Evaluation of Neural Text-to-Speech Synthesis. *arXiv* **2022**, arXiv:2204.03040.
112. Jia, Y.; Ramanovich, M.T.; Wang, Q.; Zen, H. CVSS corpus and massively multilingual speech-to-speech translation. *arXiv* **2022**, arXiv:2201.03713.
113. Shi, Y.; Bu, H.; Xu, X.; Zhang, S.; Li, M. Aishell-3: A multi-speaker mandarin tts corpus and the baselines. *arXiv* **2020**, arXiv:2010.11567.
114. Puchtler, P.; Wirth, J.; Peinl, R. Hui-audio-corpus-german: A high quality tts dataset. In Proceedings of the KI 2021: Advances in Artificial Intelligence: 44th German Conference on AI, Virtual Event, 27 September–1 October 2021; pp. 204–216.
115. Mussakhojayeva, S.; Janaliyeva, A.; Mirzakhmetov, A.; Khassanov, Y.; Varol, H.A. Kazakhtts: An open-source kazakh text-to-speech synthesis dataset. *arXiv* **2021**, arXiv:2104.08459.
116. Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F.M.; Weber, G. Common voice: A massively-multilingual speech corpus. *arXiv* **2019**, arXiv:1912.06670.
117. Harwath, D.; Recasens, A.; Surís, D.; Chuang, G.; Torralba, A.; Glass, J. Jointly discovering visual objects and spoken words from raw sensory input. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 649–665.
118. Palmer, I.; Rouditchenko, A.; Barbu, A.; Katz, B.; Glass, J. Spoken ObjectNet: A bias-controlled spoken caption dataset. *arXiv* **2021**, arXiv:2110.07575.
119. Harwath, D.; Glass, J. Deep multimodal semantic embeddings for speech and images. In Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, AZ, USA, 13–17 December 2015; pp. 237–244.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.