applied sciences

MDPI

Article

Effectively Obtaining Acoustic, Visual, and Textual Data from Videos

Jorge E. León [1,*] and Miguel Carrasco [2]

[1] Faculty of Engineering and Science, Adolfo Ibánez University (UAI), Santiago 7941169, Chile
[2] Faculty of Engineering and Science, Diego Portales University (UDP), Santiago 8370191, Chile; miguel.carrasco@udp.cl
* Correspondence: jorgleon@alumnos.uai.cl

Featured Application

The proposed method for generating large-scale audio–image–text datasets from videos addresses the critical scarcity of high-quality multimodal data, enabling advancements in audio-conditioned image-to-image generation and related tasks. By extracting semantically aligned audio–image pairs and augmenting them with descriptive texts, this work facilitates the training of more robust multimodal models, such as those for enhancing low-resolution recordings, creating dynamic video content like music videos or virtual assistant interactions, and developing augmented reality systems that incorporate real-time environmental audio for immersive user experiences. Ultimately, it promotes the democratization of AI by providing accessible, diverse datasets that support transfer learning and reduce reliance on modality conversions, paving the way for innovative applications in fields such as creative media production, remote sensing, and deep audiovisual learning.

Abstract

The increasing use of machine learning models has amplified the demand for high-quality, large-scale multimodal datasets. However, the availability of such datasets, especially those combining acoustic, visual, and textual data, remains limited. This paper addresses this gap by proposing a method of extracting related audio–image–text observations from videos. We detail the process of selecting suitable videos, extracting relevant data pairs, and generating descriptive texts using image-to-text models. Our approach ensures a robust semantic connection between modalities, enhancing the utility of the created datasets for various applications. We also explore the obtained data, discuss the challenges encountered, and propose solutions to improve data quality. The resulting datasets, which are publicly available, aim to support and advance research in multimodal data analysis and machine learning.

Keywords: data generation; multimodal data; image; audio; text; video

Academic Editor: Eui-Nam Huh

Received: 28 October 2025
Revised: 20 November 2025
Accepted: 24 November 2025
Published: 28 November 2025

Citation: León, J.E.; Carrasco, M. Effectively Obtaining Acoustic, Visual, and Textual Data from Videos. *Appl. Sci.* **2025**, *15*, 12654. https://doi.org/10.3390/app152312654

1. Introduction

In recent years, there has been an unprecedented development in the world of machine learning [1]. Several models have begun to excel in creative activities (previously considered exclusive to human minds by many) [2,3], even using non-specialized hardware [4]. In this scenario, models that can generate text associated with an image have emerged [5–7]; similarly, others that, based on texts/prompts, are capable of generating images that can fairly faithfully represent said texts have appeared [2,8–10]. An example of this can be seen in Figure 1.

From this last task, usually referred to as text-to-image, several others emerge, such as inpainting [11], outpainting [12], and image-to-image [13,14]. Commonly, the conditioning of all the aforementioned tasks tends to be text-based, and there are a few popular datasets for training such models [15–18]. This is a simple example of multimodal data being used today.

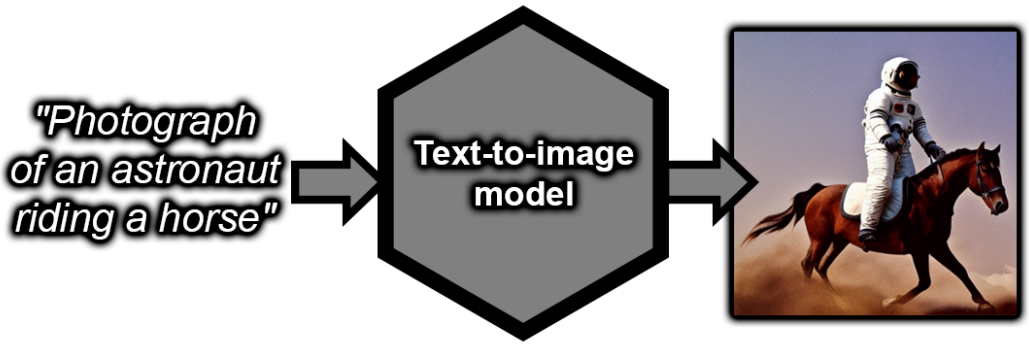

Figure 1. Text -to-image generation example. Text-to-image is a technique that generates images from textual descriptions, allowing users to create visual content based on their written prompts. Some popular models that perform this task are Stable Diffusion [8,19–21], DALL·E [22–24], Imagen [9,25], and FLUX [10].

However, it is not unheard of to find undesirable entries in any third-party dataset [26] or a lack of datasets for specific tasks (e.g., medical image analysis [27]). Similar inconveniences can also be found when dealing with datasets that include audio [28]. In particular, it has been mentioned that, relative to image datasets, audiovisual datasets are few and far between [29]. Currently, this, in turn, can be explained by the apparent low motivation for exploring fields such as audio-conditioned image-to-image [30–32], in contrast with text-conditioned image-to-image [2,8,13,19]. While there are numerous image-to-image works that condition the input image using text, there are not many that do so with audio (whether with or without added text involved), nor are there common guidelines helping researchers form these datasets on their own. Additionally, as we will explain below, the option of adapting textual/visual models to work with acoustic inputs has significant drawbacks that discourage it, rather than directly training an acoustic model for the given task.

It goes without saying that efforts in this topic could have an impact on the following: multimodal data analysis, correction of low-quality/low-resolution recordings, video generation for various purposes (virtual assistants, music videos, video transitions, etc.), democratization of artificial intelligence, augmented reality that incorporates the user's environmental audio, and transfer learning with multimodal models, among others [33–36].

In light of the above and with the desire to work with a specific type of acoustic–visual data, we formalized a method to generate audio–image–text observations based on videos (including the textual modality, in order to expand the utility of our datasets), and we employed it to generate the data that we desired for our future research. This paper delves into all of that.

In summary, in this paper, we address the need for high-quality, large-scale multimodal datasets that combine acoustic, visual, and textual data (which are currently limited). Keeping in mind the importance of maintaining a strict semantic connection between audio and visual data to improve dataset quality, as well as the ideal of minimizing data modality conversions to preserve data integrity and quality, we propose a coherent and systematic approach to extracting audio–image–text observations from videos. We discuss our results, as we generated more than 2,000,000 audio–image pairs from over 280,000 videos, together

with the transformation that we utilized to obtain the respective texts and some pending challenges that we encountered along the process.

2. Related Work

Our literature review provided clear evidence on the existence of relationships between audio and text that represent the same situation, as well as between audio and images, that should be further exploited by research and modern models (for a small summary on generative tasks that involve said modality combinations, consult Table 1).

Table 1. A summary of the most common generative audio–text and audio–image tasks.

Task	Description	Nuances
Image-to-audio	Based on an image, an audio is generated that conveys the same semantic information as the input image.	Advances have been made in the generation of audios that mimic the possible soundscape for a given image [31,37]. In a similar fashion, audios can also be generated from videos, which are nothing more than an ordered collection of images [35,36].
Text-to-audio	Based on a text, an audio is generated that conveys the same semantic information as the input text.	Some models are able to resemble a human voice reading the text given as input (subtask usually referred to as text-to-speech [34,36,38,39]). Moreover, some even make music [40] and generate lyrics based on text input [41], or they generate sounds that accommodate a given description [31,42–44].
Audio-to-image	Based on an audio, an image is generated that conveys the same semantic information as the input audio.	Voice recordings can be used to condition the modification of human faces so their mouths adapt to the corresponding sounds (i.e., lip sync [34,45]), and even the whole face can be created from scratch with the aforementioned recordings [36]. In addition, some models are capable of representing scenarios where a specific audio is produced [31,35].
Audio-to-text	Based on an audio, a text is generated that conveys the same semantic information as the input audio.	The most popular subtask here is probably speech transcription (or recognition) [34,46–48]. However, models that remarkably generate text description (or captions) from audios in general have begun to arise in recent years [31,49–51].

Exploring the cases most relevant to image-to-image conditioned by audio, there are some examples of image generation based on audio and text [52,53], and there are even cases of image-to-image generation assisted only by audio, but for mainly specific cases such as face changes (which replace a person's features with another's while maintaining consistency with the original voice recording) or lip synchronizations (where, for an image of a person, a video is generated while simulating mouth movement according to a voice recording) [34,36], which could be labeled more as a case of inpainting than image-to-image. Finally, advances in other similar areas can also be highlighted (such as text-to-video, appreciable with models such as Sora [54,55], Veo [56], Gen-3 [57], and Movie Gen [58]), and more information on some of these developments can be found at [59,60].

Currently, image-to-image generation conditioned by audio is a little-explored area of high interest in the community. To the best of our knowledge, one of the best models to date for this task is the recent CoDi model [31]. This is a model that can take any combination of audio, image, text, and video inputs and create material of any of those types (a task they called any-to-any). Additionally, a new version (CoDi-2) has also been published, and it is more flexible and adapted to conversations [30]. Another similar option is NExT-GPT, which also allows for a conversational creative process, and it works as well with audio, image, text, and video inputs [61]. Despite their promising results for future iterations, they

have not yet reached a quality that could be considered ideal. Probably, the best open-source model for this task is BindDiffusion [32]. This model is based on both the image generation model Stable Diffusion [19] and the multimodal encoder ImageBind, which incorporates six modalities, including, predictably, audio and images [62]. Notwithstanding its apparently higher quality than CoDi or NExT-GPT, it also has room for improvement, and it is not evident that it is always advisable to include the largest possible number of data modalities in these models (as seems to have been attempted in all of these cases).

The datasets involved in the training of the three previous models also shed some light on the lack of and demand for more multimodal datasets. For instance, CoDi needs to leverage different datasets (namely, LAION-400M [63], AudioSet [64], AudioCaps [65], Freesound 500K, BBC Sound Effect, SoundNet [66], WebVid-10M [67], and HD-VILA-100M [68]), with none of them combining all the required modalities. Similarly, ImageBind also makes use of multiple datasets (namely, AudioSet, SUN RGB-D [69], LLVIP [70], Ego4D [71], and private "large-scale web datasets with (image, text) pairings"), presumably due to a lack of simultaneous modalities and/or a small number of observations in each dataset. Lastly, the NExT-GPT team curated their own public dataset (called MosIT), with all the modalities they were interested in, although it only encompasses 5000 observations. We later compare the available datasets of those just mentioned with the one we generated.

3. State of the Art

In the last decade, image generation has experienced enormous growth, driven by significant advances in fields such as artificial intelligence, machine learning, and computer vision [72,73]. This progress has led to the creation of increasingly realistic and stylized images [74]. While, thanks to advances in the quality of computer-generated images (with recent examples such as Stable Diffusion XL [8] or 3 [20], DALL·E 3 [23,24], Imagen 3 [9], or FLUX [10]), the level of these images has reached a degree that makes it difficult to differentiate them from human-generated images; there is still much work to be done in terms of improving quality consistency, reducing bias, lowering computational costs, and facilitating user control over the generations (i.e., generating what the user actually expects/wants) [2].

To address this last challenge, one of the strategies that has been adopted is to increase the number of data modalities that the models receive (i.e., the types of data that are taken as input; e.g., text, image, audio, etc.) [31,46,60,75]. It is pertinent to comment that this increase in the number of modalities not only allows for greater control on the respective tasks but also opens a way to perform new ones (for example, a detailed analysis can be seen in [76], where the capabilities of GPT-4V, a colossal multimodal model of text and images, are particularly studied). In order to better illustrate the concept of data modalities, and inspired by the classification of data types explained in [45], in Figure 2, we present a conceptual map of the types of data modalities that can be used, along with examples for each (for the sake of brevity, in our conceptual map, we just include the most popular examples). An example of the use of multiple data modalities tends to be seen in image-to-image generation, where an image is taken as a reference to generate a new image, since the input image is usually accompanied by a text or a label to better condition/guide the final result [13]. In contrast, as seen in Section 2, audio-conditioned image-to-image generation has not been explored as much as text-conditioned image-to-image generation. The latter may be because working with audio is not as intuitive as working with text [77,78], but that does not invalidate the potential benefits that could be obtained by using audio in certain scenarios (such as those mentioned in Section 1).

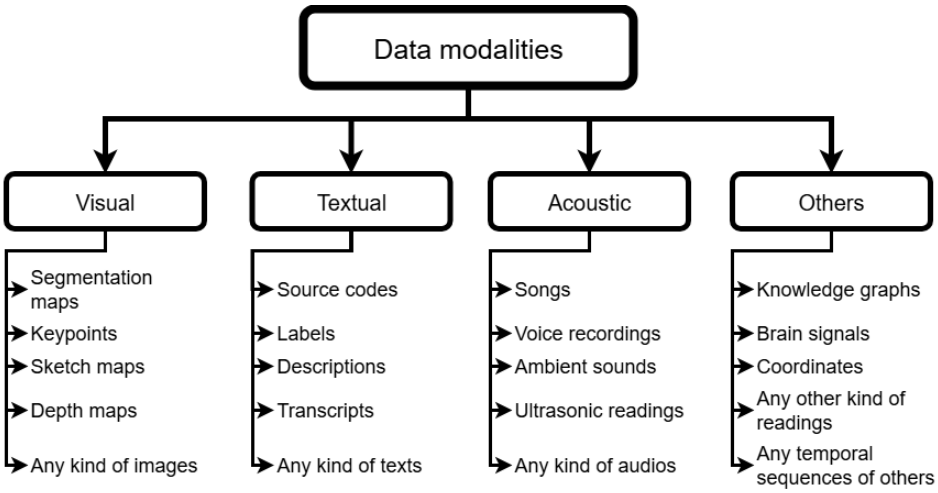

Figure 2. Types of data modalities. In [45], one can find a more detailed explanation of each modality with examples. It should be kept in mind that, in this research, we mainly care about the visual [79–82], the textual [83–85], and the acoustic [86–88] modalities.

Despite what we just wrote, we could still come up with ways to adapt the use of existing models to work with different data modalities from those that were originally intended. For instance, given the mentioned advancements in image-to-image models that are conditioned on textual inputs, it could be worth considering a new approach for scenarios where the objective is to perform image-to-image generations using audio instead of text. A logical strategy for this goal could be to transcribe the audios into the corresponding textual representations/descriptions, which could then be utilized within existing text–image models. This method should leverage the strengths of well-established text–image models, potentially validating their use with audio–image data or data of another kind that are unlike the originally intended text–image data. However, it is crucial to acknowledge that, in addition to the fact that fields such as audio-to-text conversion are still evolving and have not received as much attention as their visual counterparts [35,47,89,90], such approach presents several challenges that should be kept in mind. Let us review the main ones:

A Word limit in current models: currently, the problem of increasing the token window (i.e., words and characters) of text-to-image and audio-to-text models is open. For example, Stable Diffusion (an open-source neural network model that generates images based on text and/or image [19]) has a context window of 75 tokens [91].

B Compatibility between text–image and audio–text models: Even if a capacity of hundreds of thousands of tokens is reached to describe any audio (as can be seen analogously in certain current text generation models [92–95]), the syntax of the text obtained with such an audio–text model must match that used by the respective text–image model with which it is to be combined in order to maximize communication between the two [19,76,90,96].

C Noise incorporation (see [97] for a brief classical exploration of the definition of the term): In addition to the above, it has repeatedly been shown that transforming one modality to another is prone to incorporating noise or failing (to some extent) due to the noise that the data contain beforehand [47,98–100]. As a result, the more transformations we make, the more noise we risk adding in the process.

D Incorporation of biases: Finally, it is pertinent to highlight that, influenced both by the data and their training architectures and configurations, models tend to prioritize and specialize in certain types of audio and have their own preferences for describing them [101–104]. For example, typical cases of this can be seen in the underestima-

tion/distortion of the order of events [90,96] or in the omission of details considered irrelevant [96,105].

It is due to these reasons that, even if, in some cases, audios could/can be converted into texts and images conditioned with the generated texts, this is a significantly more problematic approach than just using audios and images. For this reason, in this research, we claim that, when working with a given set of modalities, it is convenient to perform the lowest number of data modality conversions possible. Furthermore, we believe that more audio–image research is needed to better address the respective tasks instead of just attempting to get by with what is already available.

Complementarily, it is relevant to point out that, as alluded to in [37,49–51], there are not many public datasets with audio–text pairs. In our opinion, and despite the issues enumerated previously with modality transformation approaches, the best that can be done in this scenario is to leverage a model such as CLIP [5] or BLIP [6], which, for a frame/image of a video, could return a descriptive text. Said text could, in turn, be paired with a section of audio from the video that coincides with the time interval from which the frame was extracted.

The generation of audio–image–text observations could easily be automated, so the biggest challenge would lie in finding relevant and varied videos (as well as those free from copyright conflicts). In any case, the videos collected in other research could be leveraged, within which there are recordings of musical instruments [106,107], as well as various objects, animals [108,109], and even different everyday environments [66,110].

Regarding the kind of data collected, while it would be interesting to include relationships according to the lexical meaning of spoken words, as done in [111] by relating spoken numbers and drawings of them, it would probably be better to focus on strictly non-abstract and non-artificial relationships (i.e., sounds only related to the recordings of when they were generated). This would restrict the training of the any model with these data (simplifying the range of relationships they must incorporate), facilitating convergence, and it could even make its generations more intuitive.

In summary, we have noted a valuable opportunity to explore audio–image and audio–text tasks. This demands a great volume of data, for which there are not as many datasets as one would hope for or common guidelines on how to collect the data. Due to this, in this research, we precisely propose a method of obtaining related audio–image–text observations from videos, and we describe the datasets that we created with it.

4. Materials and Methods

In the rapidly evolving landscape of multimodal data research, the integration of different kinds of data has become increasingly relevant. In this this section, we will outline a systematic approach to generate audio–image–text observations from videos. By leveraging high-quality video content, we aim to extract meaningful audio–image pairs and generate descriptive texts that enhance the utility of the resulting datasets. None of the steps that we explain below is truly novel by itself, but the combination of all of them is. We chose to unify these techniques, as they have been validated by previous works, although many did not received a proper explanation or even a brief mention. This method should be helpful to face the current scarcity of comprehensive multimodal datasets.

We will describe a multimodal data collection and processing method (specifically, for generating audio–image–text observations based on videos). It involves three key phases, which, for order's sake, we will explain in subsections of their own: 1. Initially, suitable videos are selected, prioritizing high-quality and continuous recordings, with synchronized audio and frames (see Section 4.1). 2. The next phase involves extracting audio–image pairs from these videos, ensuring that the audio is closely tied to the visual content and

minimizing file sizes without significant information loss (see Section 4.1). 3. Finally, the extracted pairs are used to generate descriptive texts using an image-to-text model, creating a comprehensive dataset for further use (see Section 4.3). All of this is also illustrated in Figure 3.

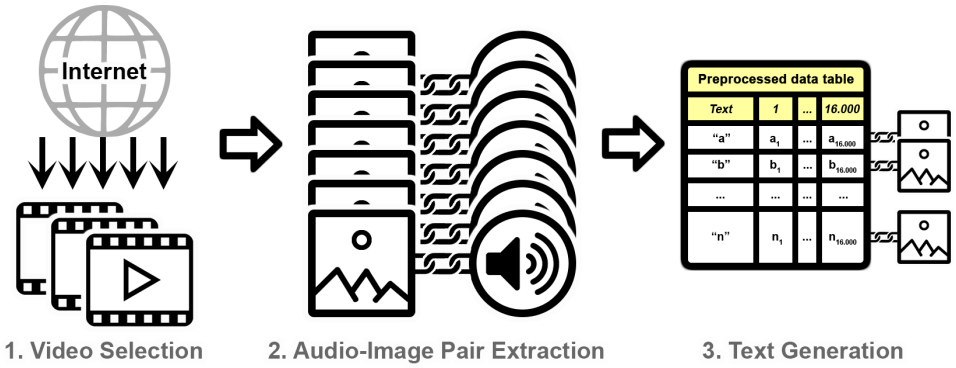

Figure 3. Summary of the whole method. 1. Video Selection: This initial phase involves identifying and selecting high-quality, continuous video recordings that feature synchronized audio and frames, ideally ensuring a strong semantic connection between the modalities (i.e., both audio and image in each pair are extracted from and related to the same situation). 2. Audio–Image Pair Extraction: In this step, audio segments are extracted from the selected videos, paired with corresponding frames, and processed to minimize file sizes while retaining essential information. 3. Text Generation: The final phase utilizes an image-to-text model to generate descriptive texts for each audio–image pair, creating a comprehensive dataset with enhanced utility.

4.1. Video Selection

First of all, it is essential to talk about the videos that we would want to work with. In addition to obviously avoiding copyrighted material and favoring HD videos with Hi-Fi audio, ideally we would hope to mainly use continuous recordings (i.e., without cuts or mixes), where each audio recording is strictly associated with the footage (i.e., without sounds that are not actually being produced in the images). Ensuring that the audio is strictly associated with the corresponding images/frames will allow for a consistent and accurate semantic connection between them, regardless of the task for which the data is being used. Additionally, using continuous recordings increases the likelihood of finding suitable video fragments to convert, especially when seeking longer audio segments.

Once again, as stated in Section 3, one can leverage public material from other research, such as that from [112]. Keep in mind that if we are not certain of the alignment between the frames and audios of our videos, we could perform an acoustic–visual alignment preprocessing step to fix this issue [113,114]. However, as we will see in the next subsection, we do not require much precision and, therefore, consider such preprocessing futile in our method. After we have collected our videos, we can start extracting pairs that consist of an image and its corresponding audio.

4.2. Audio–Image Pair Extraction

Let us define the properties of the images and audios with which we will work; they are designed to minimize their size as much as possible while preserving their core contents. Based on what has been seen in other works that generated good results [34,53,111,115], we would generally advise for the images to have a resolution of 512 $\times$ 512 pixels, in .jpg or .png format (.jpg is probably the best option, as it usually uses less space) and in RGB24 (a standard color model, consisting of a red channel, a green channel, and a blue channel, with values ranging from 0 to 255); for the audios, we would suggest a duration of 1 s, with 16,000 Hz and 16 bit depth, and they should be in .wav format and monophonic (i.e., with

a single channel). In fact, these are the properties we chose for the datasets that we will show in Section 5.

Regarding the strategy for extracting audio–image pairs, the following procedure is proposed (which is also summarized in Figure 4).

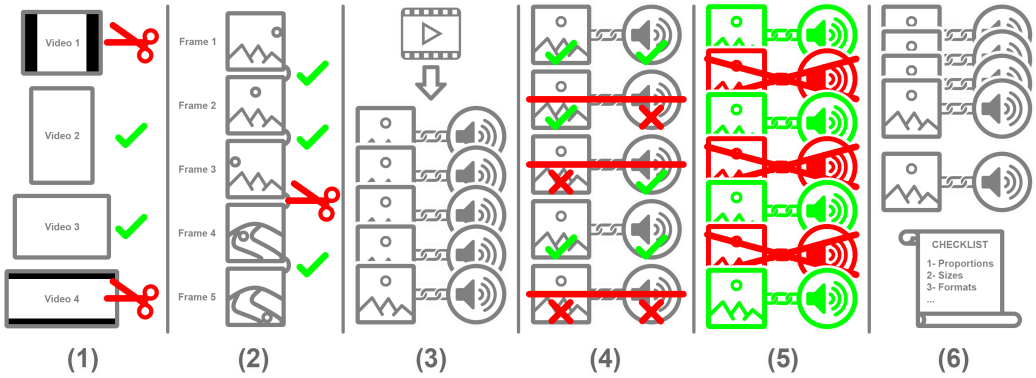

Figure 4. Summary of the audio–image pair extraction procedure. (1) Removal of black borders. (2) Discontinuous footage separation. (3) Initial audio–image pair extraction. (4) Discard of deficient audio–image pairs. (5) Skipping of pairs to enhance diversity. (6) Enforcement of the correct properties.

1. Inspect each video, evaluating if it has black borders. If so, these must be cropped to only consider relevant information in the final images. This can be accomplished by taking a frame in the middle and verifying if each first and last column/row does not have a pixel with value higher than a certain threshold in any of its channels (we use a threshold of 15 for this). In that situation, that column/row should be deleted, and the step is repeated until all of the black borders have been erased (similar to what is done in [116]).
2. To ensure that no drastic/unnatural changes are present in any audio, go through each video frame by frame. If an abrupt change is detected (for example, if the average of the squared differences of pixels between two consecutive frames is greater than a threshold of 90), then proceed to divide the footage into two and, for all purposes, treat them as distinct videos going forward. This is known as shot boundary detection [117], particularly of the pixel-based approach kind (there is no well-extended threshold for this, so it is up to the architect of the respective dataset to find an appropriate value; we empirically found that the aforementioned 90 for the average of the squared differences of pixels between two consecutive frames works for our interests, but it may be different in other cases). It should be remarked that the videos with fade transitions could present some problem with this approach and, to compensate it, more future frames could be used in the comparison.
3. For each resulting video fragment, extract consecutive audios of one second, along with the frame that is approximately in the middle of that time interval to form the respective pairs. Please note that this is known as middle frame extraction, and it is a well-extended heuristic to select a representative frame of a video fragment, which should have better odds to properly match semantically with the respective audio [118–121]. If the final part does not reach one second, it must be ignored.
4. Discard pairs whose audio has at least a given amount of continuous silence, as they will probably not contain enough information to be useful (we looked for continuous intervals of 0.5 s where none of their samples had an absolute value higher than 100) (keep in mind that we are considering samples of 16 bits, implying that the values they take go from $-32{,}768$ up to 32,767; once again, there is no well-extended threshold for this task, and the value we chose was found empirically). This, in turn, can be combined with a discount of pairs where the mean of all pixels in the image does not

surpass a given threshold (we suggest a threshold of 10). The latter should further ensure that no frames that are too dark are included.

5. To increase diversity in the data (and thus not skew the research), also consider skipping a given number of pairs from each video fragment (we just kept one pair from every three).
6. To preserve the dimensions, crop each frame according to the smaller dimension and around the center of the image, and rescale to 512 × 512 pixels. In addition, make sure to use the correct configuration for both saved files.

An example of steps 3, 4, and 5 of this process, where the extraction and filtering of audio–image pairs is used on an isolated video fragment, is shown in Figure 5. The blue fragment is discarded, as it is the last one and it does not reach a duration of 1 s. The red fragment is also not considered due to it having at least 0.5 s of continous silence. Finally, just one of every three pairs is considered (denoted by their green color) in order to increase the diversity in the data.

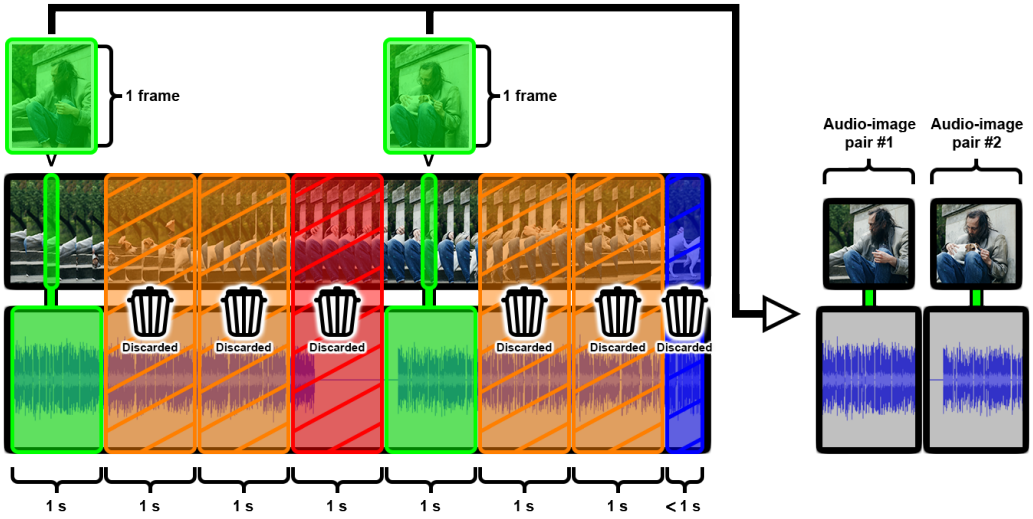

Figure 5. Application example of steps 3, 4, and 5 of the audio–image pair generation process, where the pairs are extracted from a video fragment and filtered according to our needs.

4.3. Text Generation

After we generated our audio–image pairs, we are ready to create the respective texts for each one of them. As suggested in Section 3, this will be accomplished by taking each image from every pair and, based on it, generating an associated text with the image-to-text model BLIP [6] (for the sake of speed, we used 2 beams, with a minimum length of 10 tokens and a maximum of 20); the audio will be represented as a vector of 16,000 elements, with signed 16 bit integers. The motivation for using an image-to-text model lies in the fact that the manual writing of textual descriptions for each audio–image pair is a time-consuming process that makes it impractical for a large number of observations. BLIP is also a well-tested model with a relatively fast inference time under the appropriate configuration, and it has a validated performance [122,123]. It is worth commenting that this modern possibility of leveraging image-to-text models is not something particularly novel and has also been validated in similar research [124–126]. A reasonable alternative would be to employ an audio-to-text model instead (like the ones mentioned in Table 1), although such models still need more development before being used reliably in tasks like this. In the end, both the text and the vector will then be added as a new observation/entity in a data table so that they can be manipulated more easily. For better results, it should also be pointed out that if the reader has access to a more advanced computer, then a more sophisticated image-to-text model (such as BLIP-2 [7]) should be leveraged instead of BLIP.

An illustration of how this preprocessing would look is shown in Figure 6. It is relevant to note that the use of a table to save the resulting data is an optional step, and the data can be stored in any form that best suits the user.

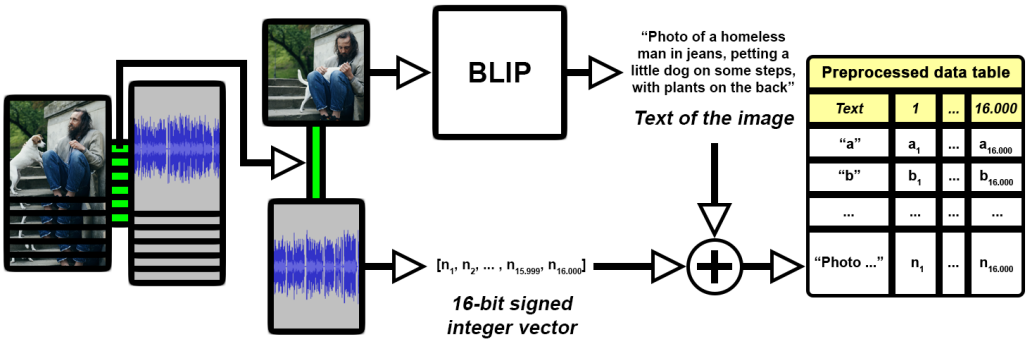

Figure 6. Example of the final data preprocessing. Audio–image pairs are expanded to include a textual modality by generating descriptive texts based on each image. These texts, along with corresponding audio values, are saved in a structured table to ease the use of the resulting dataset.

An interesting nuance to highlight is that, in the world of audio processing, there has been a tendency to prefer converting audios into spectrograms, moving from the temporal space to the temporal–spectral space, in order to facilitate pattern extraction with classical methods. This is still seen with more modern techniques [111,115,127], but, in this paper, we are only interested in creating datasets with common audio. Therefore, such a conversion is omitted in our case.

It is also worth mentioning that the minimum number of observations (or samples) "necessary" to train machine learning models is a topic open to debate (even for LLMs [128]). While a popular rule of thumb is to employ at least ten times the number of parameters of the respective model, more formal and older estimations determined that twenty times the number of parameters would be reasonable [129]. This aspect should be kept in mind when generating any dataset to train machine learning models.

Lastly, we can comment that one could even incorporate complementary data created by any-to-any models [30,31,62]. While this could be enticing at a first glance, it is a must to always remember the concerns presented in Section 3 about generating or converting data with third-party models (whether publicly validated or not). For most cases, we strongly advise prioritizing non-artificial data.

5. Results and Discussion

This section presents the datasets that we created using the method described in Section 4.

To create our audio–image pairs, we utilized videos from the public MUSIC [106,107], AudioSetZSL [108,109], and SoundNet datasets [66,110]. The videos that we employed from MUSIC dataset only contain solo performances of twenty-one different kinds of instruments, while the other two are much more diverse, ranging from musical instruments to various objects and animals, as well as different everyday environments. This diversity is relevant, as it brings substantial versatility to our final dataset. Given that AudioSetZSL is intended solely for research purposes, our datasets will also be made available for research use only, ensuring that they contribute to the advancement of multimodal data analysis while adhering to ethical standards in data sharing.

We applied the method outlined in Section 4 to process 282,081 videos, resulting in the generation of 2,240,231 audio–image pairs. From these pairs, 63,849 come from MUSIC dataset, 546,254 from AudioSetZSL, and 1,630,128 from SoundNet (proportions

that can be further appreciated in Figure 7). This predominance of observations coming from SoundNet implies no issues, as this one is the most diverse dataset of the ones we chose, and we were precisely interested in favoring it to maximize the variety in our final dataset. To exemplify the kinds of images obtained by each of these datasets, in Figure 8, we exhibit a small random sample of images, highlighting the dataset from which they derived. These audio–image pairs have been named with numbers, in a rising manner, and they are organized into 639 separate .zip files, which are publicly accessible on Kaggle, categorized into the following three distinct datasets: *Part* 1/3: https://www.kaggle.com/datasets/jorvan/image-audio-pairs-1-of-3, *Part* 2/3: https://www.kaggle.com/datasets/jorvan/image-audio-pairs-2-of-3, and *Part* 3/3: https://www.kaggle.com/datasets/jorvan/image-audio-pairs-3-of-3 (accessed on 23 November 2025). This structured approach not only facilitates easy access for researchers but also promotes further exploration and utilization of the datasets in various multimodal applications.

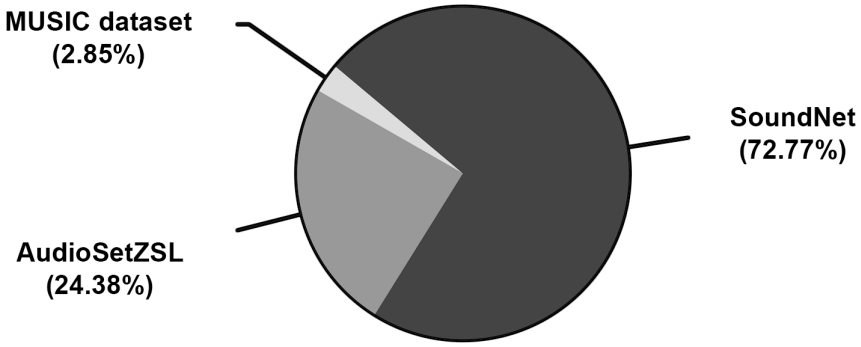

Figure 7. Visualization of the sources of our data, with the approximate percentage of observations derived from each dataset.

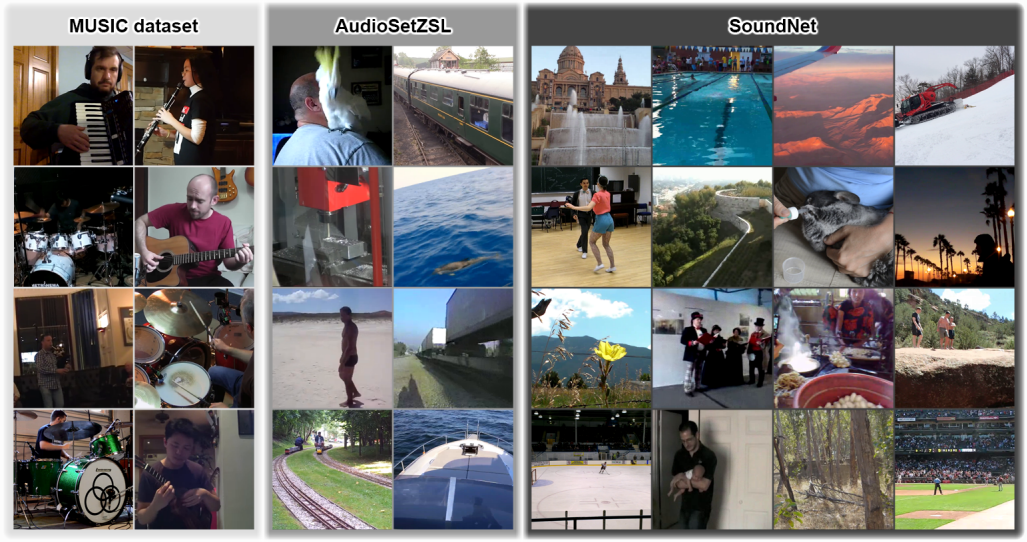

Figure 8. Sample of images derived from each of our three source datasets. As we mentioned, images coming from MUSIC dataset only contain solo performances of twenty-one different kinds of instruments, while the other two are much more diverse, ranging from musical instruments to various objects and animals, as well as different everyday environments. Additionally, all images were forced to share the same properties (particularly, a resolution of 512 × 512 pixels, .png format, and RGB24).

It is relevant to point out that some possibly problematic frames for certain uses were deemed acceptable by our filters. Namely, we noted that frames with text, with blurry images, and/or with mainly a plain color were included (see images *#817922*, *#817723* and

#1323895, respectively, from row b in Figure 9). This reinforces the value of corroborating that the original videos selected for the audio–image pair extraction process align with our interests, meaning that it would be ideal to make sure that no video with flaws that we cannot fix should be considered in the first place. Of course, curating lists of hundreds of thousands of videos is unfeasible for many researchers, which implies that the heavy work must focus on harnessing videos collected by other individuals, as well as applying the respective filters to assure the desired properties. As seen with our results, the specific filters we employed seem to still have room for improvement. The impact of these questionable frames on subsequent model training should not be underestimated, as they introduce noise into the visual modality, which can degrade model performance. Future work could improve dataset quality by adding post-processing filters for blurriness (e.g., using a Laplacian operator [130]) or for detecting and removing frames with heavy text overlays. In any case, these cases we mentioned are a minority in our data (for instance, in our aforementioned random sample of observations, we only found six possibly problematic frames, which translates to ∼10%). Nevertheless, it is important to keep this in mind and we think that the removal of these pairs could also lead to some interesting research.

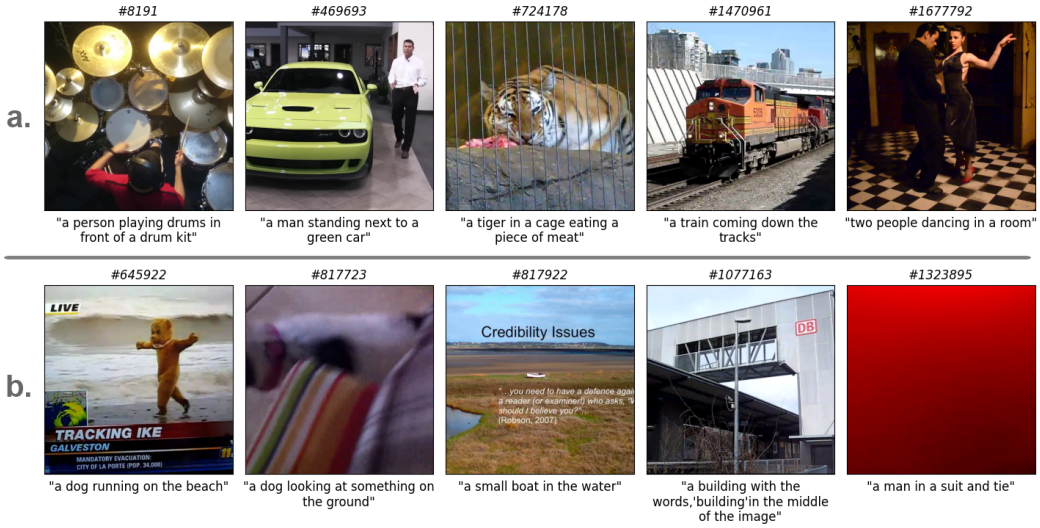

Figure 9. (**a**) A small selection of what we label sufficient and (**b**) insufficient quality image–text pairs from a random sample of 60 observations (all of them available with their respective audios at [131]). We deem *#645922* insufficient because the image has text and is from a screen; the associated text is wrong on the subject of the sentence, as it clearly shows a person in a costume and not a dog. Image *#817723* is insufficient as the image is too blurry to make a reasonable guess on what it is showing. Image *#817922* has text in the image, and the associated text is wrong. In *#1077163*, the text is also mistaken. Finally, *#1323895* has a useless image and a made-up description.

In a subsequent step, we generated descriptive captions for all of the images using the BLIP model. We paired these texts with the respective audios and stored them in 893 .csv tables. These, in turn, were saved in 263 .zip files, along with the associated images, and the numeric names were preserved. The final audio–image–text data can be found in the following 4 public datasets that we uploaded to Kaggle: *Part* 1/4: https://www.kaggle.com/datasets/jorvan/text-audio-pairs-1-of-3, *Part* 2/4: https://www.kaggle.com/datasets/jorvan/text-audio-pairs-2-of-3, *Part* 3/4: https://www.kaggle.com/datasets/jorvan/text-audio-pairs-2-of-4, and *Part* 4/4: https://www.kaggle.com/datasets/jorvan/text-audio-pairs-4-of-4 (accessed on 23 November 2025). In addition, as we cannot properly share audio through this document, we have prepared a public page, where we share 60 random samples of our final datasets to give a more solid idea of our results [131].

Going into greater detail regarding the final data, we conducted a small statistical study across all the texts. We confirmed that all descriptions have a length from 1 to 16 words, where the mean is 7.37 and the standard deviation is 1.74, approximately. We regard these values as appropriate to avoid redundancies from the image-to-text model. For comparison, we can comment that the well-known acoustic–textual dataset AudioCaps ended up with an average of 9.03 words per description [65], which does not stray too far from our result. We also counted the number of different words in the texts and found out that there were 8824 different words in use. From the list of different words, we discarded prepositions, pronouns, conjunctions, and determiners, ending up with a new total of 8652 different words. Finally, we went through the latter preprocessed list, counting the number of times that each word appeared in an observation (counting just once per observation). We show the top 60 words with the biggest percentages of presence across all observations in Table 2. From this, we can confirm that, despite a significant amount of observations containing situations featuring people, these are not the majority according to the text descriptions. Moreover, as planned, there is a nice range of diversity, given the varied collection of words that can be seen in Table 2.

Table 2. The top 60 words that appear in most observations of our final datasets (#X means the word is the Xth most common one). Prepositions, pronouns, conjunctions, and determiners are not considered, and percentages in parentheses show the proportion of observations that include them (keep in mind that observations can contain multiple or none of these words, so these percentages are not meant to add up to a hundred).

#1 people:	15.67%	#16 white:	3.84%	#31 train:	2.55%	#46 red:	1.56%
#2 man:	15.58%	#17 road:	3.68%	#32 sky:	2.48%	#47 beach:	1.55%
#3 person:	9.31%	#18 words:	3.44%	#33 building:	2.39%	#48 tree:	1.51%
#4 group:	9.14%	#19 two:	3.35%	#34 floor:	2.32%	#49 bird:	1.4%
#5 car:	8.93%	#20 driving:	3.21%	#35 dog:	2.24%	#50 guitar:	1.38%
#6 playing:	7.66%	#21 table:	3.18%	#36 middle:	2.01%	#51 shirt:	1.36%
#7 sitting:	7.65%	#22 crowd:	3.14%	#37 cars:	1.94%	#52 boat:	1.35%
#8 room:	7.25%	#23 suit:	3.13%	#38 holding:	1.77%	#53 parking:	1.35%
#9 street:	6.83%	#24 field:	2.99%	#39 child:	1.7%	#54 little:	1.27%
#10 background:	6.6%	#25 water:	2.97%	#40 trees:	1.69%	#55 band:	1.27%
#11 down:	5.8%	#26 front:	2.96%	#41 riding:	1.67%	#56 girl:	1.25%
#12 standing:	4.68%	#27 city:	2.93%	#42 cat:	1.67%	#57 truck:	1.25%
#13 woman:	4.58%	#28 tie:	2.87%	#43 laying:	1.66%	#58 bed:	1.23%
#14 walking:	4.44%	#29 parked:	2.73%	#44 black:	1.61%	#59 chair:	1.19%
#15 baby:	3.93%	#30 stage:	2.66%	#45 night:	1.56%	#60 wall:	1.16%

To attest to the usefulness of our data, we also conducted two additional tests to inspect both biases and diversity in our audios. On one hand, for the biases, we created a 65,536 $\times$ 16,000-matrix (coinciding with our chosen bit depth and total samples per audio, respectively), filled with zeros, and we proceeded to add to each element the count of times where the corresponding instantaneous amplitude was present in the given timestamps across all audios. We then plotted the resulting matrix (assigning to zero the white color and to the maximum count of the matrix a the black color, with all the counts in between a gray color that linearly denotes its closeness to each extreme) and obtained the result, as shown on the right side of Figure 10, which also illustrates the whole procedure. The observable Gaussian distributions in all timestamps coincide neatly with the theory in [132,133], and thus, this shows that no evident biases are present. Despite this, we also analyzed the skew and kurtosis of each timestamp. We found that the skews are approximately equal to -0.04 ± 0.02, while the kurtoses are 9.93 ± 0.09, which means that, due to the high kurtoses, these are not real normal distributions [134]. Regardless, situations like this tend to be seen when working with large samples of values (as this makes the kurtosis more sensitive to outliers [135]), and, more importantly, this fact has no major implications that we may be

concerned about. As an added point, we can also consider that such a leptokurtic (or super-Gaussian) distribution is a well-documented characteristic of speech and audio signals, which have a sharp peak at zero (representing the noise floor), with fat tails (representing sparse high-amplitude events) [136]. On the other hand, for the diversity, we calculated the corresponding Acoustic Diversity Index (ADI) [137–140]. This is a popular metric based on the Shannon index [141], and it has high popularity for audio diversity measurements (especially in fields related to biology, although it can be employed with any kind of audio dataset). We summarize its calculation in Figure 11, and we also need to point out that, in our case, this metric may take values between 0 and ∼3.4657 (with bigger values conveying a greater diversity). Dividing this interval into three equally distributed ones, we end up with the following: $[0, 1.1552]$ for low diversity values, $(1.1552, 2.3105)$ for medium diversity values, and $[2.3105, 3.4657]$ for high diversity values. Our resulting ADI was ∼3.0525, which serves as further evidence of the diversity in our dataset.

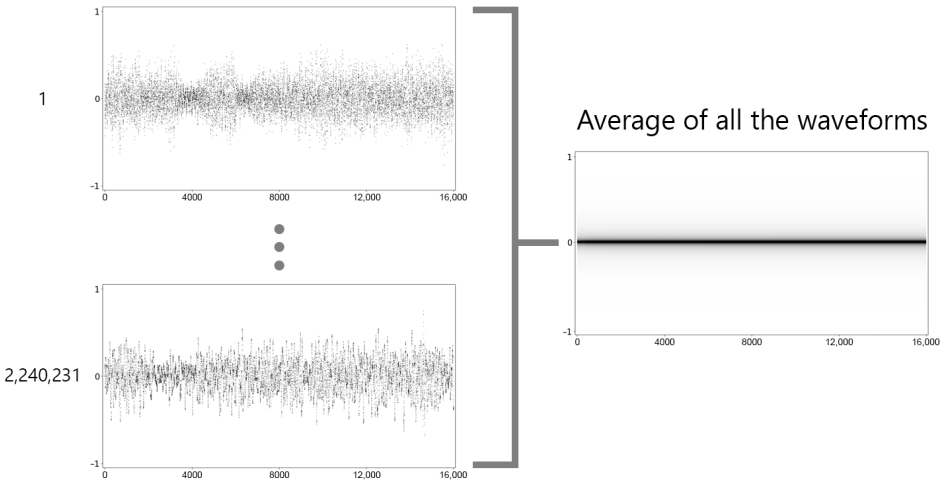

Figure 10. Average of all waveforms in our observations. The horizontal axis contains the timestamps, while the vertical one is for the instantaneous amplitudes. Take into consideration that, analogous to the images, all the audios share the same properties (these being a duration of 1 s, with 16,000 Hz, a bit depth of 16, and a monophonic channel).

Complementarily, the contribution of our dataset becomes clear in Table 3, where we compare the scope of our dataset with those of the datasets mentioned in Section 2 (leveraged by models that include acoustic, textual, and visual modalities), as well as some additional datasets that could also be employed in similar audio–image–text tasks. As we can see, most datasets do not even contain one million observations, which is a real handicap, given that modern models deal with millions of parameters and, therefore, require larger datasets to be properly trained. Currently, researchers need to arduously search for multiple datasets and artfully come up with ways to utilize them in audio–image–text tasks, as they not only be too small but also do not contain all the modalities needed and/or their contents are too specialized (not to mention the extra preprocessing steps one must add when the data are not homogeneous). All of this hinders the potential research that could be done in the field, and thus, we expect both our dataset and our detailed method contribute to easing this struggle, especially when noticing the large supply of audiovisual datasets.

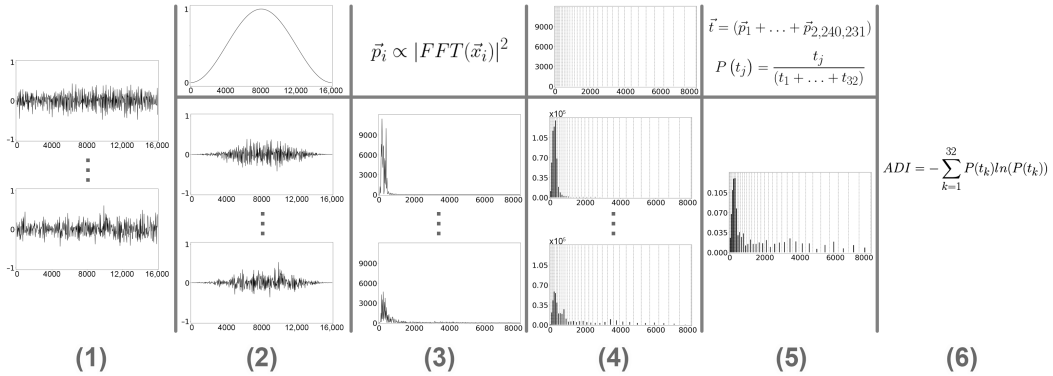

Figure 11. Summary of our ADI calculation. (1) We take all of our audios in their raw form. (2) We apply the Hann function [142] (visible on the top) over each audio signal, so all of them loop smoothly, and we avoid spectral leakage. (3) We compute the fast Fourier transform of each signal from the previous step, and we get the magnitude of each resulting complex number and square it; now these new values are proportional to the real power spectra, and we can treat these as substitutes for them in the next steps. (4) We generate 32 evenly spaced bins on the mel scale [143] of our range of frequencies (i.e., $[0, 8000]$) and aggregate the respective values that share each bin. (5) We sum all of our 2,240,231 vectors of grouped power spectra, preserving their bins and dividing each resulting component by the sum of all of them combined; this effectively leaves us with the probabilities of presence of each interval of frequencies in our audios. (6) Finally, we apply the Shannon index [141] over our probabilities of the previous step in order to obtain our ADI (we adhered to the common practices of using the ADI metric, which include the use of the mel scale and a small number of bins [137–140]).

Once again, there are two shortcomings that we must highlight. The first one is that, naturally, there may be a semantic drift in which a generated text accurately describes the respective image but is irrelevant to the concurrent audio. This potential disconnection between modalities may bring some issues, so it cements the need to perfect our method with future tools (possibly, adding an audio-to-text model into the pipeline). The second one is that, despite the relatively long time taken to create the text descriptions, the BLIP configuration used was fairly basic in maximizing speed. This means that the text quality is not nearly as high as one would wish for in some instances. To illustrate the latter, let us look at Figure 9. Contrasting with the appropriate descriptions that we get in cases such as row a, row b presents various kinds of errors. Image *#645922* misidentifies a person in a costume as a dog, *#817723* has an imprecise caption due to the poor resolution of the image, *#817922* may also be negatively affected by the presence of texts, *#1077163* imagines a text that does not exist, and *#1323895* hallucinates the presence of a man when it is clearly just a color gradient. Again, the quality of the images we work with has a fundamental influence on the quality of our final texts, but so does the model we use. For other researchers, we strongly recommend the use of better hardware, as well as a better image-to-text model than we used (to run BLIP, we only had a MacBook Pro available (with a, M1 chip, 8 cores, and 8 Gb of unified memory)). As a final proposal, the use of multiple image-to-text models could be considered, possibly even including audio-to-text ones. The outputs of these models could be fed into a large language model, which can generate a new text that averages and encompasses the semantic meaning all descriptions, improving the chances of ending up with an appropriate caption.

Table 3. A comparison table between many multimodal datasets and ours. **A** means that the observations include **Audios**, **I** means the same for **Images**, **T** for **Texts**, and **V** for **Videos**. ✔ means that the data modality is present in the respective dataset. K stands for thousands, and M stands for millions.

Name of the Dataset	A	I	T	V	# of Samples	Contents
AudioCaps [65]	✔		✔		>45.5K	Alarms, various objects and animals, natural phenomena, and different everyday environments.
AudioSet [64]	✔		✔		>2.0M	632 audio event classes, including musical instruments, various objects and animals, and different everyday environments.
CMU-MOSEI [144]	✔	✔	✔		>3.2K	People speaking directly to a camera in monolog form, intended for sentiment analysis.
Ego4D [71]	✔			✔	>5.8K	Egocentric video footage of different everyday situations, with portions of the videos accompanied by audio and/or 3D meshes of the environment.
Flickr30k Entities [145]		✔	✔		>31.7K	Diverse environments, objects, animals, and activities, with the addition to bounding boxes to the image–text pairs.
Freesound 500K [31]	✔		✔		500.0K	Diverse situations, sampled from the Freesound website and accompanied by tags and descriptions.
HD-VILA-100M [68]			✔	✔	>100.0M	A wide range of categories, including tutorials, vlogs, sports, events, animals, and films.
InternVid [146]			✔	✔	>233.0M	Diverse environments, objects, animals, activities, and everyday situations.
LAION-400M [63]		✔	✔		400.0M	Everyday scenes, animals, activities, art, scientific imagery, and various objects.
LLVIP [70]		✔			>16.8K	Street environments, where each visible-light image is paired with an infrared one of the same scene.
MMIS [147]	✔	✔	✔		>150.0K	A wide range of interior spaces, capturing various styles, layouts, and furnishings.
MosIT [61]	✔	✔	✔	✔	5.0K	Diverse environments, objects, animals, artistic elements, activities, and conversations.
SUN RGB-D [69]		✔			>10.0K	Everyday environments, where each image has the depth information of the various objects in it.
SoundNet [66]	✔			✔	>2.1M	Videos without professional edition, depicting natural environments, everyday situations, and various objects and animals.
WebVid-10M [67]			✔	✔	>10.0M	Natural environments, everyday situations, and various objects and animals.
AVT Multimodal Dataset (Ours)	✔	✔	✔		>2.2M	Musical instruments, various objects and animals, and different everyday environments.

6. Conclusions

In this study, we tackled the significant challenge of generating high-quality multimodal datasets, specifically focusing on audio–image–text observations derived from videos. Our motivation stemmed from the increasing demand for diverse and large-scale datasets in the machine learning community, particularly for multimodal data that include audio, which are often scarce [29,37].

We proposed a method of generating these datasets by leveraging continuous video recordings, ensuring a strict semantic connection between acoustic and visual data (i.e., both audio and image in each pair are extracted from and related to the same situation). This

approach addresses the common issue of undesirable entries in third-party datasets [26] and the lack of datasets for specific tasks, such as medical image analysis [27], reinforcement learning [148], and audio–text in general [37,49–51].

Our method involved three key steps: collecting suitable videos, extracting audio–image pairs, and generating textual descriptions for each pair using the BLIP model [6]. This process resulted in the creation of over 2 million audio–image pairs, which were further extended to include textual descriptions, forming a comprehensive multimodal dataset. Despite some limitations, such as the inclusion of frames with text or blurry images, as well as the basic configuration used for text generation, our dataset represents an advancement in the availability of multimodal data for research purposes.

The literature review highlighted the potential of exploiting relationships between audio, image, and text data [33,35,45,60]. These relationships can enhance various applications, including multimodal data analysis, correction of low-quality recordings, video generation, augmented reality, and transfer learning with multimodal models.

Our research underscores the importance of minimizing data modality conversions to preserve data quality. We also emphasized the need for more research on audio–image and audio–text tasks given the current lack of high-quality data and guidelines for new researchers.

Future work could focus on refining the filtering process to exclude undesirable frames more effectively, ensuring temporal alignment by incorporating more recent techniques in the pipeline (such as those seen in [113,149,150]), employing more advanced image-to-text models to improve the quality of textual descriptions, and even leveraging future audio-to-text models, which could complement the aforementioned descriptions. Additionally, exploring the potential of incorporating complementary data from any-to-any models while being mindful of the concerns related to data modality conversions could further enhance the utility of these datasets.

Overall, our contributions provide a valuable resource for the research community and highlight the importance of multimodal data in advancing machine learning models. The impact of this method and our resulting dataset extends beyond audio-conditioned image generation. We believe that this resource can significantly benefit other multimodal tasks, such as cross-modal retrieval (e.g., searching for images using audio queries), acoustic–visual source separation, and generative video content creation. We hope that our work will inspire further research and development in this area, ultimately leading to more robust and versatile AI systems.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The final audio–image–text data can be found in the following four public datasets that we uploaded to Kaggle: *Part* 1/4: https://www.kaggle.com/datasets/jorvan/text-audio-pairs-1-of-3, *Part* 2/4: https://www.kaggle.com/datasets/jorvan/text-audio-pairs-2-of-3, *Part* 3/4: https://www.kaggle.com/datasets/jorvan/text-audio-pairs-2-of-4, and *Part* 4/4: https://www.kaggle.com/datasets/jorvan/text-audio-pairs-4-of-4 (accessed on 23 November 2025).

Acknowledgments: During the preparation of this manuscript/study, the authors used Stable Diffusion 1.5 for the purposes of generating the example in Figure 1. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

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

References

1. Mukhamediev, R.I.; Symagulov, A.; Kuchin, Y.; Yakunin, K.; Yelis, M. From Classical Machine Learning to Deep Neural Networks: A Simplified Scientometric Review. *Appl. Sci.* **2021**, *11*, 5541. [CrossRef]
2. Zhang, C.; Zhang, C.; Zhang, M.; Kweon, I.S. Text-to-image Diffusion Models in Generative AI: A Survey. *arXiv* **2023**, arXiv:2303.07909.
3. Franceschelli, G.; Musolesi, M. Creativity and Machine Learning: A Survey. *arXiv* **2022**, arXiv:2104.02726. [CrossRef]
4. Dhar, S.; Guo, J.; Liu, J.J.; Tripathi, S.; Kurup, U.; Shah, M. A Survey of On-Device Machine Learning: An Algorithms and Learning Theory Perspective. *ACM Trans. Internet Things* **2021**, *2*, 15. [CrossRef]
5. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models from Natural Language Supervision. *arXiv* **2021**, arXiv:2103.00020. [CrossRef]
6. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *arXiv* **2022**, arXiv:2201.12086.
7. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv* **2023**, arXiv:2301.12597.
8. Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; Rombach, R. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. *arXiv* **2023**, arXiv:2307.01952. [CrossRef]
9. Imagen-Team-Google; Baldridge, J.; Bauer, J.; Bhutani, M.; Brichtova, N.; Bunner, A.; Castrejon, L.; Chan, K.; Chen, Y.; Dieleman, S.; et al. Imagen 3. *arXiv* **2024**, arXiv:2408.07009. [CrossRef]
10. Labs, B.F. FLUX. 2024. Available online: https://github.com/black-forest-labs/flux (accessed on 23 November 2025).
11. Bertalmío, M.; Sapiro, G.; Caselles, V.; Ballester, C. Image inpainting. In Proceedings of the 27th Internationl Conference on Computer Graphics and Interactive Techniques Conference, New Orleans, LA, USA, 23–28 July 2000; pp. 417–424.
12. Singh, S.; Aggarwal, N.; Jain, U.; Jaiswal, H. Outpainting Images and Videos using GANs. *Int. J. Comput. Trends Technol.* **2020**, *68*, 24–29. [CrossRef]
13. Pang, Y.; Lin, J.; Qin, T.; Chen, Z. Image-to-Image Translation: Methods and Applications. *IEEE Trans. Multimed.* **2022**, *24*, 3859–3881. [CrossRef]
14. Saxena, S.; Teli, M.N. Comparison and Analysis of Image-to-Image Generative Adversarial Networks: A Survey. *arXiv* **2022**, arXiv:2112.12625.
15. 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 13th European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 740–755.
16. Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In Proceedings of the 35th International Conference on Neural Information Processing Systems, Online, 6–14 December 2022; pp. 25278–25294.
17. Changpinyo, S.; Sharma, P.K.; Ding, N.; Soricut, R. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. In Proceedings of the 2021 IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 3557–3567.
18. Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; Tang, X. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1096–1104.
19. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv* **2022**, arXiv:2112.10752. [CrossRef]
20. Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. *arXiv* **2024**, arXiv:2403.03206. [CrossRef]
21. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. Stable Diffusion. 2021. Available online: https://github.com/CompVis/stable-diffusion (accessed on 23 November 2025).
22. Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-Shot Text-to-Image Generation. *arXiv* **2021**, arXiv:2102.12092.
23. Betker, J.; Goh, G.; Jing, L.; Brooks, T.; Wang, J.; Li, L.; Ouyang, L.; Zhuang, J.; Lee, J.; Guo, Y.; et al. Improving Image Generation with Better Captions. 2023. Available online: https://cdn.openai.com/papers/dall-e-3.pdf (accessed on 23 November 2025).

24. OpenAI. DALL·E 3 System Card. 2023. Available online: https://cdn.openai.com/papers/DALL_E_3_System_Card.pdf (accessed on 23 November 2025).
25. Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E.; Ghasemipour, S.K.S.; Ayan, B.K.; Mahdavi, S.S.; Gontijo-Lopes, R.; et al. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2024; pp. 36479–36494.
26. Birhane, A.; Prabhu, V. Large image datasets: A pyrrhic win for computer vision? In Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 1536–1546.
27. Shorten, C.; Khoshgoftaar, T. A survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
28. Wijngaard, G.; Formisano, E.; Esposito, M.; Dumontier, M. Audio-Language Datasets of Scenes and Events: A Survey. *arXiv* **2024**, arXiv:2407.06947. [CrossRef]
29. Żelaszczyk, M.; Mańdziuk, J. Audio-to-Image Cross-Modal Generation. *arXiv* **2021**, arXiv:2109.13354.
30. Tang, Z.; Yang, Z.; Khademi, M.; Liu, Y.; Zhu, C.; Bansal, M. CoDi-2: In-Context, Interleaved, and Interactive Any-to-Any Generation. *arXiv* **2023**, arXiv:2311.18775.
31. Tang, Z.; Yang, Z.; Zhu, C.; Zeng, M.; Bansal, M. Any-to-any generation via composable diffusion. In Proceedings of the 37th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 10–15 December 2024; pp. 16083–16099.
32. Lab, S.A. BindDiffusion: One Diffusion Model to Bind Them All. 2024. Available online: https://github.com/sail-sg/BindDiffusion (accessed on 23 November 2025).
33. Zheng, Z.; Chen, J.; Zheng, X.; Lu, X. Remote Sensing Image Generation From Audio. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 994–998. [CrossRef]
34. Khanjani, Z.; Watson, G.; Janeja, V.P. Audio deepfakes: A survey. *Front. Big Data* **2023**, *5*, 1001063. [CrossRef]
35. Zhu, H.; Luo, M.D.; Wang, R.; Zheng, A.H.; He, R. Deep Audio-visual Learning: A Survey. *Int. J. Autom. Comput.* **2021**, *18*, 351–376. [CrossRef]
36. Shi, Z. A Survey on Audio Synthesis and Audio-Visual Multimodal Processing. *arXiv* **2021**, arXiv:2108.00443. [CrossRef]
37. Sheffer, R.; Adi, Y. I Hear Your True Colors: Image Guided Audio Generation. *arXiv* **2023**, arXiv:2211.03089. [CrossRef]
38. Wang, C.; Chen, S.; Wu, Y.; Zhang, Z.; Zhou, L.; Liu, S.; Chen, Z.; Liu, Y.; Wang, H.; Li, J.; et al. Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. *arXiv* **2023**, arXiv:2301.02111. [CrossRef]
39. Wu, H.; Chen, X.; Lin, Y.C.; Chang, K.-w.; Chung, H.L.; Liu, A.H.; Li, H.-y. Towards audio language modeling—An overview. *arXiv* **2024**, arXiv:2402.13236.
40. Melechovsky, J.; Guo, Z.; Ghosal, D.; Majumder, N.; Herremans, D.; Poria, S. Mustango: Toward Controllable Text-to-Music Generation. In Proceedings of the 2024 North American Chapter of the Association for Computational Linguistics, Mexico City, Mexico, 16–21 June 2024; pp. 8293–8316.
41. Dhariwal, P.; Jun, H.; Payne, C.; Kim, J.W.; Radford, A.; Sutskever, I. Jukebox: A Generative Model for Music. *arXiv* **2020**, arXiv:2005.00341. [CrossRef]
42. Valle, R.; Badlani, R.; Kong, Z.; Lee, S.-g.; Goel, A.; Kim, S.; Santos, J.F.; Dai, S.; Gururani, S.; AIJa'fari, A.; et al. Fugatto 1—Foundational Generative Audio Transformer Opus 1. 2024. Available online: https://openreview.net/forum?id=B2Fqu7Y2cd (accessed on 23 November 2025).
43. Kreuk, F.; Synnaeve, G.; Polyak, A.; Singer, U.; Défossez, A.; Copet, J.; Parikh, D.; Taigman, Y.; Adi, Y. AudioGen: Textually Guided Audio Generation. *arXiv* **2023**, arXiv:2209.15352. [CrossRef]
44. Liu, H.; Chen, Z.; Yuan, Y.; Mei, X.; Liu, X.; Mandic, D.; Wang, W.; Plumbley, M.D. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 21450–21474.
45. Zhan, F.; Yu, Y.; Wu, R.; Zhang, J.; Lu, S.; Liu, L.; Kortylewski, A.; Theobalt, C.; Xing, E. Multimodal Image Synthesis and Editing: The Generative AI Era. *arXiv* **2023**, arXiv:2112.13592. [CrossRef]
46. Xu, P.; Zhu, X.; Clifton, D.A. Multimodal Learning with Transformers: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12113–12132. [CrossRef]
47. Ansari, F.; Gupta, R.; Singh, U.; Shaikh, F. Transcripter-Generation of the transcript from audio to text using Deep Learning. *Int. J. Comput. Sci. Eng.* **2019**, *7*, 770–773. [CrossRef]
48. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; pp. 28492–28518.
49. Bai, J.; Liu, H.; Wang, M.; Shi, D.; Plumbley, M.; Gan, W.S.; Chen, J. AudioSetCaps: An Enriched Audio-Caption Dataset using Automated Generation Pipeline with Large Audio and Language Models. In Proceedings of the Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation, Vancouver, BC, Canada, 9–15 December 2024.

50. Lanzendörfer, L.A.; Pinkl, C.; Perraudin, N.; Wattenhofer, R. BLAP: Bootstrapping Language-Audio Pre-training for Music Captioning. In Proceedings of the Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation, Vancouver, BC, Canada, 9–15 December 2024.
51. Xu, X.; Zhang, Z.; Zhou, Z.; Zhang, P.; Xie, Z.; Wu, M.; Zhu, K.Q. BLAT: Bootstrapping Language-Audio Pre-training based on AudioSet Tag-guided Synthetic Data. In Proceedings of the 31st ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 2756–2764.
52. Yariv, G.; Gat, I.; Wolf, L.; Adi, Y.; Schwartz, I. AudioToken: Adaptation of Text-Conditioned Diffusion Models for Audio-to-Image Generation. *arXiv* **2023**, arXiv:2305.13050.
53. Jonason, N.; Sturm, B.L.T. TimbreCLIP: Connecting Timbre to Text and Images. *arXiv* **2022**, arXiv:2211.11225. [CrossRef]
54. Liu, Y.; Zhang, K.; Li, Y.; Yan, Z.; Gao, C.; Chen, R.; Yuan, Z.; Huang, Y.; Sun, H.; Gao, J.; et al. Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models. *arXiv* **2024**, arXiv:2402.17177. [CrossRef]
55. OpenAI. Video Generation Models as World Simulators. 2024. Available online: https://openai.com/index/video-generation-models-as-world-simulators/ (accessed on 23 November 2025).
56. DeepMind, G. Veo. 2024. Available online: https://deepmind.google/technologies/veo/ (accessed on 23 November 2025).
57. Runway. Introducing Gen-3 Alpha: A New Frontier for Video Generation. 2024. Available online: https://runwayml.com/research/introducing-gen-3-alpha (accessed on 23 November 2025).
58. The Movie Gen Team. Movie Gen: A Cast of Media Foundation Models. 2024. Available online: https://ai.meta.com/static-resource/movie-gen-research-paper (accessed on 23 November 2025).
59. Cao, H.; Tan, C.; Gao, Z.; Xu, Y.; Chen, G.; Heng, P.A.; Li, S.Z. A Survey on Generative Diffusion Models. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 2814–2830. [CrossRef]
60. Suzuki, M.; Matsuo, Y. A survey of multimodal deep generative models. *Adv. Robot.* **2022**, *36*, 261–278. [CrossRef]
61. Wu, S.; Fei, H.; Qu, L.; Ji, W.; Chua, T.S. NExT-GPT: Any-to-Any Multimodal LLM. *arXiv* **2024**, arXiv:2309.05519.
62. Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K.V.; Joulin, A.; Misra, I. ImageBind: One Embedding Space To Bind Them All. *arXiv* **2023**, arXiv:2305.05665. [CrossRef]
63. Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; Komatsuzaki, A. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *arXiv* **2021**, arXiv:2111.02114.
64. Gemmeke, J.F.; Ellis, D.P.W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio Set: An ontology and human-labeled dataset for audio events. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing, New Orleans, LA, USA, 5–9 March 2017; pp. 776–780.
65. Kim, C.D.; Kim, B.; Lee, H.; Kim, G. AudioCaps: Generating Captions for Audios in The Wild. In Proceedings of the 2019 North American Chapter of the Association for Computational Linguistics, New Minneapolis, MN, USA, 3–5 June 2019; pp. 119–132.
66. Aytar, Y.; Vondrick, C.; Torralba, A. SoundNet: Learning Sound Representations from Unlabeled Video. In Proceedings of the 30th International Conference on Neural Information Processing Systems, Changsha, China, 20–23 November 2016; pp. 892–900.
67. Bain, M.; Nagrani, A.; Varol, G.; Zisserman, A. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In Proceedings of the 2021 IEEE International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 1708–1718.
68. Xue, H.; Hang, T.; Zeng, Y.; Sun, Y.; Liu, B.; Yang, H.; Fu, J.; Guo, B. Advancing High-Resolution Video-Language Representation with Large-Scale Video Transcriptions. In Proceedings of the 2022 IEEE International Conference on Computer Vision, New Orleans, LA, USA, 18–24 June 2022; pp. 5026–5035.
69. Song, S.; Lichtenberg, S.P.; Xiao, J. SUN RGB-D: A RGB-D scene understanding benchmark suite. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 567–576.
70. Jia, X.; Zhu, C.; Li, M.; Tang, W.; Zhou, W. LLVIP: A Visible-infrared Paired Dataset for Low-light Vision. In Proceedings of the 2021 IEEE International Conference on Computer Vision Workshops, Montreal, BC, Canada, 11–17 October 2021; pp. 3489–3497.
71. Grauman, K.; Westbury, A.; Byrne, E.; Cartillier, V.; Chavis, Z.; Furnari, A.; Girdhar, R.; Hamburger, J.; Jiang, H.; Kukreja, D.; et al. Ego4D: Around the World in 3000 Hours of Egocentric Video. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, 1–32. [CrossRef]
72. Bie, F.; Yang, Y.; Zhou, Z.; Ghanem, A.; Zhang, M.; Yao, Z.; Wu, X.; Holmes, C.; Golnari, P.; Clifton, D.A.; et al. RenAIssance: A Survey into AI Text-to-Image Generation in the Era of Large Model. *arXiv* **2023**, arXiv:2309.00810. [CrossRef]
73. Gao, L.; Biderman, S.; Black, S.; Golding, L.; Hoppe, T.; Foster, C.; Phang, J.; He, H.; Thite, A.; Nabeshima, N.; et al. The Pile: An 800GB Dataset of Diverse Text for Language Modeling. *arXiv* **2020**, arXiv:2101.00027. [CrossRef]
74. Elasri, M.; Elharrouss, O.; Al-Maadeed, S.; Tairi, H. Image Generation: A Review. *Neural Process. Lett.* **2022**, *54*, 4609–4646. [CrossRef]
75. Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; Lillicrap, T.; baptiste Alayrac, J.; Soricut, R.; Lazaridou, A.; Firat, O.; Schrittwieser, J.; et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv* **2024**, arXiv:2403.05530. [CrossRef]

76. Yang, Z.; Li, L.; Lin, K.; Wang, J.; Lin, C.C.; Liu, Z.; Wang, L. The Dawn of LMMs: Preliminary Explorations with GPT-4V(ision). *arXiv* **2023**, arXiv:2309.17421. [CrossRef]
77. Agostinelli, A.; Denk, T.I.; Borsos, Z.; Engel, J.; Verzetti, M.; Caillon, A.; Huang, Q.; Jansen, A.; Roberts, A.; Tagliasacchi, M.; et al. MusicLM: Generating Music From Text. *arXiv* **2023**, arXiv:2301.11325. [CrossRef]
78. Hong, J.; Park, S.; Ro, Y. Intuitive Multilingual Audio-Visual Speech Recognition with a Single-Trained Model. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 4886–4890.
79. Trigka, M.; Dritsas, E. A Comprehensive Survey of Deep Learning Approaches in Image Processing. *Sensors* **2025**, *25*, 531. [CrossRef] [PubMed]
80. Sinha, R.K.; Pandey, R.; Pattnaik, R. Deep Learning For Computer Vision Tasks: A review. *arXiv* **2018**, arXiv:1804.03928. [CrossRef]
81. Guo, Y.; Liu, Y.; Oerlemans, A.; Lao, S.; Wu, S.; Lew, M.S. Deep learning for visual understanding: A review. *Neurocomputing* **2016**, *187*, 27–48. [CrossRef]
82. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep Learning for Computer Vision: A Brief Review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [CrossRef]
83. Otter, D.W.; Medina, J.R.; Kalita, J.K. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 604–624. [CrossRef]
84. Bai, Y.; Wang, D.Z. More Than Reading Comprehension: A Survey on Datasets and Metrics of Textual Question Answering. *arXiv* **2022**, arXiv:2109.12264. [CrossRef]
85. Taha, K.; Yoo, P.D.; Yeun, C.; Homouz, D.; Taha, A. A comprehensive survey of text classification techniques and their research applications: Observational and experimental insights. *Comput. Sci. Rev.* **2024**, *54*, 100664. [CrossRef]
86. Bianco, M.J.; Gerstoft, P.; Traer, J.; Ozanich, E.; Roch, M.A.; Gannot, S.; Deledalle, C.A. Machine learning in acoustics: Theory and applications. *J. Acoust. Soc. Am.* **2019**, *146*, 3590–3628. [CrossRef] [PubMed]
87. Roger, V.; Farinas, J.; Pinquier, J. Deep neural networks for automatic speech processing: A survey from large corpora to limited data. *EURASIP J. Audio Speech Music Process.* **2022**, *2022*, 19. [CrossRef]
88. Chen, X.; Chang, L.; Yu, X.; Huang, Y.; Tu, X. A Survey on World Models Grounded in Acoustic Physical Information. *arXiv* **2025**, arXiv:2506.13833. [CrossRef]
89. Vercauteren, G.; Reviers, N. Audio Describing Sound—What Sounds are Described and How?: Results from a Flemish case study. *J. Audiov. Transl.* **2022**, *5*, 114–133. [CrossRef]
90. Wu, H.H.; Nieto, O.; Bello, J.P.; Salamon, J. Audio-Text Models Do Not Yet Leverage Natural Language. In Proceedings of the 2023 IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodos, Greece, 4–10 June 2023; pp. 1–5.
91. Maks-s. Stable Diffusion Akashic Records. 2023. Available online: https://github.com/Maks-s/sd-akashic (accessed on 23 November 2025).
92. Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. The Llama 3 Herd of Models. *arXiv* **2024**, arXiv:2407.21783. [CrossRef]
93. Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv* **2024**, arXiv:2312.00752.
94. Mistral AI. Mistral Models. 2024. Available online: https://mistral.ai/technology/#models (accessed on 23 November 2025).
95. Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku. 2024. Available online: https://anthropic.com/claude-3-model-card (accessed on 23 November 2025).
96. Koo, R.; Lee, M.; Raheja, V.; Park, J.I.; Kim, Z.M.; Kang, D. Benchmarking Cognitive Biases in Large Language Models as Evaluators. *arXiv* **2023**, arXiv:2309.17012. [CrossRef]
97. Scales, J.; Snieder, R. What is noise? *Geophysics* **1998**, *63*, 1122–1124. [CrossRef]
98. Huang, R.; Long, Y.; Han, J.; Xu, H.; Liang, X.; Xu, C.; Liang, X. NLIP: Noise-Robust Language-Image Pre-training. In Proceedings of the 37th AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; pp. 926–934.
99. Yang, S.; Xu, Z.; Wang, K.; You, Y.; Yao, H.; Liu, T.; Xu, M. BiCro: Noisy Correspondence Rectification for Multi-modality Data via Bi-directional Cross-modal Similarity Consistency. In Proceedings of the 2023 IEEE Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19883–19892.
100. Kang, W.; Mun, J.; Lee, S.; Roh, B. Noise-Aware Learning from Web-Crawled Image-Text Data for Image Captioning. In Proceedings of the 2023 IEEE International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 2942–2952.
101. Belém, C.G.; Seshadri, P.; Razeghi, Y.; Singh, S. Are Models Biased on Text without Gender-related Language? In Proceedings of the 12th International Conference on Learning Representations, Vienna, Austria, 7–11 May 2024.
102. Lin, A.; Paes, L.M.; Tanneru, S.H.; Srinivas, S.; Lakkaraju, H. Word-Level Explanations for Analyzing Bias in Text-to-Image Models. *arXiv* **2023**, arXiv:2306.05500.
103. Zieliński, S.; Rumsey, F.; Bech, S. On Some Biases Encountered in Modern Audio Quality Listening Tests—A Review. *J. Audio Eng. Soc.* **2008**, *56*, 427–451.

104. Agrawal, A.; Batra, D.; Parikh, D.; Kembhavi, A. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4971–4980.
105. Qiao, T.; Zhang, J.; Xu, D.; Tao, D. MirrorGAN: Learning Text-To-Image Generation by Redescription. In Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1505–1514.
106. Zhao, H.; Gan, C.; Rouditchenko, A.; Vondrick, C.; McDermott, J.; Torralba, A. The Sound of Pixels. In Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 587–604.
107. Zhao, H.; Rouditchenko, A. MUSIC Dataset from Sound of Pixels. 2018. Available online: https://github.com/roudimit/MUSIC_dataset (accessed on 23 November 2025).
108. Parida, K.K.; Matiyali, N.; Guha, T.; Sharma, G. Coordinated Joint Multimodal Embeddings for Generalized Audio-Visual Zero-shot Classification and Retrieval of Videos. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 3240–3249.
109. Parida, K.K. AudioSetZSL. 2019. Available online: https://github.com/krantiparida/AudioSetZSL (accessed on 23 November 2025).
110. Thomee, B.; Shamma, D.A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; Li, L.J. YFCC100M: The new data in multimedia research. *Commun. ACM* **2016**, *59*, 64–73. [CrossRef]
111. Shim, J.Y.; Kim, J.; Kim, J.K. Audio-to-Visual Cross-Modal Generation of Birds. *IEEE Access* **2023**, *11*, 27719–27729. [CrossRef]
112. Zhang, Y.; Asketorp, J.; Feng, D. Awesome-Video-Datasets. 2023. Available online: https://github.com/xiaobai1217/Awesome-Video-Datasets (accessed on 23 November 2025).
113. Wang, J.; Fang, Z.; Zhao, H. AlignNet: A Unifying Approach to Audio-Visual Alignment. In Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March 2020; pp. 3298–3306.
114. Sun, J.; Deng, L.; Afouras, T.; Owens, A.; Davis, A. Eventfulness for Interactive Video Alignment. *ACM Trans. Graphs* **2023**, *42*, 46. [CrossRef]
115. Gao, R.; Oh, T.H.; Grauman, K.; Torresani, L. Listen to Look: Action Recognition by Previewing Audio. *arXiv* **2020**, arXiv:1912.04487. [CrossRef]
116. Chen, Z.; Li, B.; Ma, T.; Liu, L.; Liu, M.; Zhang, Y.; Li, G.; Li, X.; Zhou, S.; He, Q.; et al. Phantom-Data: Towards a General Subject-Consistent Video Generation Dataset. *arXiv* **2025**, arXiv:2506.18851.
117. Abdulhussain, S.H.; Ramli, A.R.; Saripan, M.I.; Mahmmod, B.M.; Al-Haddad, S.A.R.; Jassim, W.A. Methods and Challenges in Shot Boundary Detection: A Review. *Entropy* **2018**, *20*, 214. [CrossRef]
118. Lindgren, G. A Comparison Between KeyFrame Extraction Methods for Clothing Recognition. 2023. Available online: https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-507294 (accessed on 23 November 2025).
119. Hou, J.; Su, L.; Zhao, Y. Key Frame Selection for Temporal Graph Optimization of Skeleton-Based Action Recognition. *Appl. Sci.* **2024**, *14*, 9947. [CrossRef]
120. Salehi, M.; Park, J.S.; Yadav, T.; Kusupati, A.; Krishna, R.; Choi, Y.; Hajishirzi, H.; Farhadi, A. ActionAtlas: A VideoQA Benchmark for Domain-specialized Action Recognition. In Proceedings of the 37th International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 10–15 December 2024; Volume 37, pp. 137372–137402.
121. Zhang, Y.; Tokmakov, P.; Hebert, M.; Schmid, C. A Structured Model for Action Detection. In Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9967–9976.
122. Waheed, S.; An, N.M. Image Embedding Sampling Method for Diverse Captioning. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Suzhou, China, 5–9 November 2025; pp. 3141–3157.
123. Li, Q.; Xie, Y.; Grundlingh, N.; Chawan, V.R.; Wang, C. Assessing Image-Captioning Models: A Novel Framework Integrating Statistical Analysis and Metric Patterns. In Proceedings of the 7th Workshop on e-eommerce and NLP, Torino, Italy, 21 May 2024; pp. 79–87.
124. Li, H.; Xu, M.; Zhan, Y.; Mu, S.; Li, J.; Cheng, K.; Chen, Y.; Chen, T.; Ye, M.; Wang, J.; et al. OpenHumanVid: A Large-Scale High-Quality Dataset for Enhancing Human-Centric Video Generation. In Proceedings of the 2025 IEEE Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 10–17 June 2025; pp. 7752–7762.
125. Ventura, L.; Schmid, C.; Varol, G. Learning Text-to-Video Retrieval from Image Captioning. *Int. J. Comput. Vis.* **2024**, *133*, 1834–1854. [CrossRef]
126. Xue, Z.; Zhang, J.; Hu, T.; He, H.; Chen, Y.; Cai, Y.; Wang, Y.; Wang, C.; Liu, Y.; Li, X.; et al. UltraVideo: High-Quality UHD Video Dataset with Comprehensive Captions. *arXiv* **2025**, arXiv:2506.13691.
127. Wu, H.H.; Seetharaman, P.; Kumar, K.; Bello, J.P. Wav2CLIP: Learning Robust Audio Representations from Clip. In Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, 23–27 May 2022; pp. 4563–4567.
128. Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; de Las Casas, D.; Hendricks, L.A.; Welbl, J.; Clark, A.; et al. Training Compute-Optimal Large Language Models. In Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2024; pp. 30016–30030.

129. Jackson, D.L. Revisiting Sample Size and Number of Parameter Estimates: Some Support for the N:q Hypothesis. *Struct. Equ. Model. Multidiscip. J.* **2003**, *10*, 128–141. [CrossRef]
130. Bansal, R.; Raj, G.; Choudhury, T. Blur image detection using Laplacian operator and Open-CV. In Proceedings of the 2016 International Conference System Modeling & Advancement in Research Trends, Moradabad, India, 25–27 November 2016; pp. 63–67.
131. León, J.E. AVT Multimodal Dataset. 2024. Available online: https://jorvan758.github.io/AVT-Multimodal-Dataset/ (accessed on 23 November 2025).
132. Prasad, R. Does mixing of speech signals comply with central limit theorem? *Int. J. Electron. Commun.* **2008**, *62*, 782–785. [CrossRef]
133. Dehay, D.; Leskow, J.; Napolitano, A. Central Limit Theorem in the Functional Approach. *IEEE Trans. Signal Process.* **2013**, *61*, 4025–4037. [CrossRef]
134. Hatem, G.; Zeidan, J.; Goossens, M.M.; Moreira, C. Normality Testing Methods and the Importance of Skewness and Kurtosis in Statistical Analysis. *Sci. Technol.* **2022**, *3*, 7. [CrossRef]
135. Van Zyl, J.M.; Van der Merwe, S. An empirical study of the behaviour of the sample kurtosis in samples from symmetric stable distributions. *S. Afr. Stat. J.* **2020**, *54*, 255–265. [CrossRef]
136. LeBlanc, J.; De Leon, P. Speech separation by kurtosis maximization. In Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, Seattle, WA, USA, 15 May 1998; Volume 2, pp. 1029–1032.
137. Xiang, Y.; Meng, Q.; Zhang, X.; Li, M.; Yang, D.; Wu, Y. Soundscape diversity: Evaluation indices of the sound environment in urban green spaces—Effectiveness, role, and interpretation. *Ecol. Indic.* **2023**, *154*, 110725. [CrossRef]
138. Bradfer-Lawrence, T.; Desjonqueres, C.; Eldridge, A.; Johnston, A.; Metcalf, O. Using acoustic indices in ecology: Guidance on study design, analyses and interpretation. *Methods Ecol. Evol.* **2023**, *14*, 2192–2204. [CrossRef]
139. Pijanowski, B.C.; Villanueva-Rivera, L.J.; Dumyahn, S.L.; Farina, A.; Krause, B.L.; Napoletano, B.M.; Gage, S.H.; Pieretti, N. Soundscape Ecology: The Science of Sound in the Landscape. *BioScience* **2011**, *61*, 203–216. [CrossRef]
140. Bradfer-Lawrence, T.; Duthie, B.; Abrahams, C.; Adam, M.; Barnett, R.J.; Beeston, A.; Darby, J.; Dell, B.; Gardner, N.; Gasc, A.; et al. The Acoustic Index User's Guide: A practical manual for defining, generating and understanding current and future acoustic indices. *Methods Ecol. Evol.* **2025**, *16*, 1040–1050. [CrossRef]
141. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
142. Blackman, R.B.; Tukey, J.W. The measurement of power spectra from the point of view of communications engineering—Part I. *Bell Syst. Tech. J.* **1958**, *37*, 185–282. [CrossRef]
143. O'Shaughnessy, D. *Speech Communications: Human and Machine*; Wiley-IEEE Press: Hoboken, NJ, USA, 2000; Volume 2, p. 128.
144. Bagher Zadeh, A.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Melbourne, Australia, 15–20 July 2018; pp. 2236–2246.
145. Plummer, B.A.; Wang, L.; Cervantes, C.M.; Caicedo, J.C.; Hockenmaier, J.; Lazebnik, S. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2641–2649.
146. Wang, Y.; He, Y.; Li, Y.; Li, K.; Yu, J.; Ma, X.; Li, X.; Chen, G.; Chen, X.; Wang, Y.; et al. InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Generation. *arXiv* **2024**, arXiv:2307.06942.
147. Kassab, H.; Mahmoud, A.; Bahaa, M.; Mohamed, A.; Hamdi, A. MMIS: Multimodal Dataset for Interior Scene Visual Generation and Recognition. *arXiv* **2024**, arXiv:2407.05980. [CrossRef]
148. Mazzaglia, P.; Verbelen, T.; Dhoedt, B.; Courville, A.; Rajeswar, S. GenRL: Multimodal-foundation world models for generalization in embodied agents. *arXiv* **2024**, arXiv:2406.18043.
149. Han, T.; Xie, W.; Zisserman, A. Temporal Alignment Networks for Long-term Video. *arXiv* **2022**, arXiv:2204.02968. [CrossRef]
150. Viertola, I.; Iashin, V.; Rahtu, E. Temporally Aligned Audio for Video with Autoregression. In Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing, Hyderabad, India, 6–11 April 2025; pp. 1–5.