TODD A. ROBERTS (SBN 129722)
KEVIN W. ISAACSON (SBN 281067)
ROPERS MAJESKI PC
333 W. Santa Clara St., Suite 930
San Jose, CA 95113
Telephone: 408.287.6262
Facsimile: 408.918.4501
Email: todd.roberts@ropers.com
kevin.isaacson@ropers.com

Attorneys for Plaintiff
S.A. JAMENDO, a company formed under the laws of Grand Duchy of Luxembourg

UNITED STATES DISTRICT COURT

NORTHERN DISTRICT OF CALIFORNIA

S.A. JAMENDO, a company formed under the laws of Grand Duchy of Luxembourg,

Plaintiff,

v.

NVIDIA CORPORATION, a corporation organized under the laws of Delaware, and DOES 1 through 50,

Defendants.

Case No.

COMPLAINT FOR COPYRIGHT INFRINGEMENT, VIOLATION OF LICENSE AGREEMENT, UNJUST ENRICHMENT

Plaintiff S.A. JAMENDO ("Jamendo"), by and through its undersigned attorneys, for its Complaint against Defendant NVIDIA CORPORATION ("Nvidia"), hereby alleges as follows:

NATURE OF ACTION

1. Jamendo is a worldwide leader in connecting independent musicians with music lovers. It pioneered the use of permissive licensing, enabling artists to legally share their music through its platform, Jamendo.com. Through sustained effort, innovation, and consistent success over the years, Jamendo has evolved into a major platform, now hosting more than 500,000 tracks from over 40,000 independent artists worldwide.

2. Jamendo carefully curates its extensive music portfolio, enhancing accessibility through creative categorization and detailed tagging that guide users toward discovering new artists and tracks. Through these sustained curation efforts, Jamendo has developed, and is the author and owner of, a valuable copyrighted compilation of data ("Jamendo Database"), encompassing hundreds of thousands of tracks on its platform. Jamendo's original selection, coordination, and arrangement of musical works and sound recordings and associated metadata, including distinctive and granular tagging of each track.

3. Through continuous investment in refining and expanding the Jamendo Database, Jamendo strengthens the quality and relevance of its platform, attracting both artists and listeners. These ongoing efforts create a dynamic user experience and reinforce the platform's value by driving engagement, discovery, and sustained user interest.

4. Jamendo's platform also advances the commercial interest of its artists. Through its dedicated service, Jamendo Licensing, the company actively licenses tracks from its platform for commercial use. This ensures that artists not only share their music widely but also have meaningful opportunities to generate revenue from their work. For many artists and tracks, Jamendo has been granted the exclusive right to the commercial use of tracks. In such instances, artists are looking solely to Jamendo to promote and regulate the commercial use of their tracks for the mutual benefit of Jamendo and the artist.

5. By leveraging both its platform and the Jamendo Database, Jamendo fosters user engagement and deepens connections with music, which in turn drives demand for commercial licensing. This ecosystem creates a direct pathway from discovery to monetization, enabling the use of tracks in films, television shows, advertising, and other commercial applications, while ensuring that artists benefit from the growing exposure and demand for their music.

6. The Jamendo Database is a valuable asset independent of the tracks hosted on the

Jamendo platform, serving as a key catalyst for connecting users with artists. To facilitate access to music, Jamendo has, in certain limited instances, made portions of the Jamendo Database available for non-commercial research under specific agreements. However, any commercial use has always required an appropriate license from Jamendo.

7. Nvidia promotes itself as a leader in accelerated computing, built around a singular mission: solving complex challenges that others cannot. It highlights Nvidia's work in artificial intelligence (AI) and digital twin technologies, emphasizing how its innovations are transforming major global industries and driving meaningful societal impact.

8. While pursuing commercial success, Nvidia has blatantly disregarded Jamendo's intellectual property rights by using the Jamendo Database as well as the musical works and sound recordings listed in **Exhibit A** to develop its AI platforms, including training its AI engine, without any authorization from Jamendo. The musical works and sound recordings listed in Exhibit A are identified herein as a preliminary and representative sample; upon information and belief, they do not constitute an exhaustive list of registered works contained within the Jamendo Database. Jamendo is in the process of conducting a comprehensive search of the U.S. Copyright Office records, and anticipates identifying additional registered works within the Jamendo Database during the discovery process in this matter.

9. Nvidia refused to engage in meaningful discussions with Jamendo regarding its unlawful infringement when contacted by representatives of Jamendo. Nvidia has thus necessitated this action for copyright infringement, breach of contract, unjust enrichment, and unfair business practices arising from its unauthorized use of the Jamendo Database as well as the musical works and sound recordings listed in **Exhibit A** in the development, training, and operation of a commercial generative AI systems.

10. Nvidia's conduct is a blatant violation of the Copyright Act of 1976. The

Copyright Act of 1976 reflects a carefully calibrated balance between the rights of authors and the interests of the public, rooted in the constitutional mandate to "promote the Progress of Science and useful Arts" by granting authors and inventors exclusive rights to their works for limited times. *Universal City Studios, Inc. v. Sony Corp.*, 659 F.2d 963 (1981), *Threshold Media Corp. v. Relativity Media, LLC*, 166 F. Supp. 3d 1011 (2013), *Laws v. Sony Music Entm't, Inc.*, 448 F.3d 1134 (2006). This balance is achieved through a framework of exclusive rights, limitations, and exceptions designed to incentivize creativity while ensuring public access to knowledge and culture.

11. Nvidia's conduct has disregarded the balance struck by the Copyright Act of 1976 by taking and commercially using Jamendo's Dataset and many tracks for which Jamendo holds the exclusive right to use commercially. Nvidia's conduct has caused significant harm to Jamendo, including loss of control over its compilation, diminution in the value of its licensing market, and loss of revenue.

12. Jamendo seeks injunctive relief, damages, restitution, disgorgement, and all other appropriate relief.

PARTIES

13. Plaintiff Jamendo S.A. is a company formed under the laws of the Grand Duchy of Luxembourg, with its registered office and principal place of business located at 51, Rue de Strasbourg, L-2561 Luxembourg.

14. Defendant Nvidia Corporation is a corporation organized under the laws of the State of Delaware with its principal place of business located at 2788 San Tomas Expressway, Santa Clara, California 95051.

JURISDICTION AND VENUE

15. This is a civil action seeking injunctive relief and damages for copyright

infringement arising under the Copyright Act of 1976, 17 U.S.C. §§ 101, *et. seq*. This Court has subject matter jurisdiction over the claims asserted herein pursuant to 28 U.S.C. §§ 1331, 1338(a), and § 1367.

16. This Court has personal jurisdiction over Nvidia because Nvidia is a resident of this District. Nvidia maintains its principal place of business at 2788 San Tomas Expressway, Santa Clara, California 95051, and is therefore "at home" in this District.

17. Venue is proper in this District pursuant to 28 U.S.C. §§ 1391(b)(1) and 1400(a) because Nvidia resides in this District, may be found in this District, and has its principal place of business in Santa Clara, California.

STATEMENT OF FACTS

Jamendo's Business

18. Jamendo is "all about connecting musicians and music lovers from all over the world." It promotes independent artists by providing a platform where their music can be discovered for free by users for personal, non-commercial use, while also offering opportunities for commercialization through licensing for commercial purposes. Through this dual-purpose model, Jamendo has grown to host over 500,000 tracks from more than 40,000 artists across over 150 countries worldwide.

19. Jamendo's effort to promote artists begins with its unique licensing model. Jamendo requires all artists listing tracks through Jamendo to grant specific limited licenses allowing royalty-free personal use of tracks. This allows Jamendo's users to freely explore, discover, and enjoy new music. Should users wish to use tracks from Jamendo's platform beyond personal use, moving into any form of commercial, more than 65.000 emerging artists have authorized Jamendo to license commercial use of their tracks. Frequently, artists have authorized Jamendo the exclusive right to use tracks commercially, including the exclusive right to authorize

third parties' commercial use of tracks. By giving a fair share of the revenues to the artists, Jamendo helps them focus on creating music. In fact, when a license is purchased from Jamendo, the author receives up to 60% of the license price.

20. To support its dual-purpose model, Jamendo prioritizes the curation, organization, and licensing of all musical content on its platform. In doing so, it has developed and maintains internal metadata systems, including tagging frameworks, classification methods, and organizational schemas used to structure its content. Central to these efforts is the Jamendo Database, which includes metadata associated with each track. This metadata is curated through human intervention and is created and continuously updated based on editorial judgment and qualitative assessment. Such curation includes the characterization of each track's genre, mood, instrumentation, and other defining attributes.

21. The structured Jamendo Database comprises an organized compilation of metadata and content that distinguishes it from a collection of unorganized factual information. Its organization and content are specifically designed to support discovery, recommendation, and licensing use cases. In this way, the Jamendo Database serves as a catalyst for connecting users to artists through discovery and recommendation, while also connecting artists to users through the commercial licensing of tracks.

Creation of the MTG-Jamendo Database

22. The Jamendo Database is a valuable asset independent of the tracks hosted on the Jamendo platform. To facilitate access to the music of independent artists, Jamendo has, at times, made portions the Jamendo Database available for non-commercial research purposes.

23. In one instance, in or around 2019, Jamendo created a custom subset of the Jamendo Database in connection with a publicly funded academic research collaboration with the Music Technology Group ("MTG") at Universitat Pompeu Fabra in Barcelona, Spain. This subset

incorporated Jamendo's structured metadata systems and organizational schema and is known as the "MTG-Jamendo Dataset." The MTG-Jamendo Dataset was created from the larger Jamendo Database using specific creative selections of tracks that Jamendo identified as tracks that would further the research purposes of MTG.

24. The resulting MTG-Jamendo Dataset reflects and embodies Jamendo's underlying organizational structure, including its selection, coordination, and arrangement of metadata and associated content.

25. The MTG-Jamendo Dataset has been made publicly available, including through repositories hosted on GitHub, an online platform widely used to host and distribute datasets for research purposes. The MTG-Jamendo Dataset repository is available at https://github.com/MTG/mtg-jamendo-dataset.

26. The MTG-Jamendo Dataset was developed for and has been widely used in academic and research contexts, including for the evaluation and development of music-related technologies.

27. Notably, MTG describes the MTG-Jamendo Dataset as a "new open dataset for music auto-tagging" and confirms that it "contains over 55,000 full audio tracks with 195 tags from genre, instrument, and mood/theme categories." Through its repository on GitHub, MTG further states that it "provides elaborated data splits for researchers" and reports "the performance of a simple baseline approach on five different sets of tags: genre, instrument, mood/theme, top-50, and overall." The repository also "contains metadata, scripts, instructions on how to download and use the dataset and reproduce baseline results."

28. The MTG-Jamendo Dataset GitHub repository is intended for research use in research and includes a recommended citation: Bogdanov, D., Won M., Tovstogan P., Porter A., & Serra X. (2019). *The MTG-Jamendo Dataset for Automatic Music Tagging*. Machine Learning

for Music Discovery Workshop, International Conference on Machine Learning (ICML 2019).

Licensing Restrictions Governing the MTG-Jamendo Dataset

29. The GitHub repository hosting the MTG-Jamendo Dataset is publicly available but expressly states that "the metadata is licensed under a CC BY-NC-SA 4.0" license. The CC BY-NC-SA 4.0 refers to the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License, which permits use for non-commercial purposes while requiring attribution and the sharing of derivative works under the same terms. This licensing model enables researchers to use the MTG-Jamendo Dataset to advance their research objectives.

30. However, the CC BY-NC 4.0 license permits use of the MTG-Jamendo Dataset for non-commercial purposes while prohibiting use directed toward commercial advantage or monetary compensation.

31. The MTG-Jamendo Dataset repository on GitHub also includes an express notice stating that it is made available solely for non-commercial research and academic use. Specifically, the repository provides:

"Please note that the MTG-Jamendo Dataset is made available solely for non-commercial research and academic use. Any other use, including but not limited to commercial applications, requires prior written authorization from Jamendo S.A. For any commercial use requests, please contact Jamendo's licensing team at hello@jamendo.com to obtain a license adapted to your project. Commercial use is defined as any use that is intended for or directed toward commercial advantage or monetary compensation, including but not limited to uses generating revenue through advertising, affiliate programs, or product integration."

32. These notices further specify that commercial use includes any use intended for or directed toward commercial advantage or monetary compensation, including uses that generate revenue through advertising, affiliate programs, or product integration.

33. The GitHub repository through which the MTG-Jamendo Dataset is distributed provides direct access to the dataset files together with accompanying documentation and licensing information, including the CC BY-NC 4.0 license and the non-commercial use notice. These materials are presented in connection with access to and use of the MTG-Jamendo Dataset, such that users downloading and using the dataset are placed on notice of, and proceed subject to, those terms.

34. These licensing terms reflect the existence of a commercial market for licensing datasets and structured metadata for use in AI and related technologies.

Commercial Licensing Market and Value of the Jamendo Database and the MTG-Jamendo Dataset

35. Jamendo has entered into commercial agreements involving the use of the Jamendo Database, including its catalogue, metadata, and related datasets, such as the MTG-Jamendo Dataset, for AI and machine learning purposes.

36. Jamendo has been approached by multiple companies seeking access to its catalogue, metadata, and related datasets for AI-related uses. Pricing for such uses is negotiated on a case-by-case basis depending on factors including the customer profile, intended use, amount of content, territory, and duration.

37. Jamendo has not granted licenses for use directly comparable to the large-scale commercial training of generative AI systems, such as Fugatto and Audio Flamingo, without a specific negotiated agreement and compensation structure.

38. The MTG-Jamendo Dataset alone contains proprietary information related to approximately 55,000 to 56,000 tracks from approximately 3,500 unique artists.

39. A dataset of this scale and organization provides immediate access to a large catalogue of rights-cleared music with detailed metadata, classifications, and tags suitable for use

in AI training.

40. Building or licensing a comparable dataset would require substantial time, expense, and resources, including obtaining rights from thousands of artists, organizing and maintaining a large music database, manually reviewing and validating metadata, creating and refining tags and classifications, and ensuring that tracks are legally cleared and commercially usable.

41. By using the MTG-Jamendo Dataset without authorization, Nvidia obtained substantial value, including immediate access to a large, organized, and commercially useful music catalogue. Nvidia also avoided the time, cost, and effort required to independently source, clean, organize, and validate comparable data, including the value inherent in the dataset's structured organization and curated metadata.

Jamendo's Investment and Distinctive Features

42. Jamendo has invested significant resources over many years in developing and maintaining its music catalogue, platform, metadata, and related systems, including the Jamendo Database.

43. This investment includes, among other things, the technical development and maintenance of its database and platform, personnel dedicated to moderation, tagging, and curation, internal quality control and rights verification processes, and the creation and maintenance of proprietary classification and tagging systems.

44. Jamendo's employees manually review, confirm, and, where appropriate, adjust and improve metadata associated with tracks in its catalogue and the Jamendo Database. This manual review includes, among other things, genre, mood, theme, and instrument tags, as well as overall metadata accuracy and completeness.

45. Jamendo's employees curate tracks into themed playlists and commercial

categories. On average, approximately 120 tracks are manually reviewed and curated each day as part of these processes.

46. The MTG-Jamendo Dataset is widely used in the academic and research community because it combines a large catalogue of music with detailed metadata, genre tags, and audio descriptors.

47. The MTG-Jamendo Dataset represents only a subset of the broader Jamendo Database, which is significantly larger and contains additional curated information, classifications, and metadata created and maintained by Jamendo. The MTG-Jamendo Dataset is a derivative work based on the Jamendo Database.

48. What distinguishes the MTG-Jamendo Dataset is not only the number of tracks, but also the manual review and enhancement of metadata, the organization and classification of tracks, and the curated playlists and commercial categories.

49. These features reflect substantial investment, expertise, and ongoing effort, and contribute to the commercial value of the MTG-Jamendo Dataset and Jamendo's broader Jamendo Database.

Nvidia's Unauthorized Use of the MTG-Jamendo Dataset, Jamendo Database, and Tracks Exclusively Licensed to Jamendo

50. Nvidia emphasizes the transformative role of AI in shaping how people create and experience digital content. Its CEO, Jensen Huang, has stated that "AI is no longer a single breakthrough or application – it is essential infrastructure," underscoring Nvidia's view that AI underpins modern computing and creative systems. He further asserts that AI will revolutionize every industry, reflecting Nvidia's positioning of its technologies as enabling broad-based innovation, including in creative and generative domains. Huang has also described the technological shift as foundational, noting that the future is about accelerated computing and AI

working hand-in-hand, highlighting Nvidia's focus on enabling rapid, high-performance creation and deployment of AI-driven outputs.

51. Nvidia emphasizes the importance of scale, speed, and rapid innovation in the development of its AI technologies, including its generative audio models. Nvidia highlights its ability to train large-scale models on massive datasets and to accelerate the development and deployment of AI systems.

52. Nvidia has published papers describing its dataset construction and training methodologies, including *Fugatto: Foundational Generative Audio Transformer Opus 1* and *Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities*. Attached hereto as **Exhibits B - C** are true and correct copies of Nvidia's publicly-available research publications.

53. Nvidia has described Fugatto as "[its] first step toward a future where unsupervised multitask learning in audio synthesis and transformation emerges from data and model scale" and characterizes it as a "Swiss Army knife for sound," emphasizing its role in advancing generative audio capabilities.

54. Nvidia has developed Audio Flamingo as a "fully open-source Large Audio Language Model (LALM)" with "state-of-the-art performance in audio understanding and reasoning." Nvidia explains that Audio Flamingo enables "audio understanding and reasoning across speech, sound, and music," and includes capabilities such as "multi-turn, multi-audio chat" and "long-context audio reasoning," allowing the system to interpret, analyze, and reason over complex audio inputs.

55. As with any generative AI service, Nvidia must "train" its software models using large volumes of data, including sound recordings and related metadata. This data is incorporated into a training dataset that enables the system to identify patterns and generate outputs in response

to user inputs. The training process necessarily involves ingesting and copying data, as well as storing either the original data or derivative forms of it. Where copyrighted works are involved, such reproduction and the creation of derivative works must be authorized by the rights holder or otherwise constitute unlawful infringement.

56. Nvidia admits that its models are trained using "a large text and audio dataset with at least 20 million rows" and that this dataset is "exclusively comprised of open source data."

57. Nvidia further admits that it aggregates data from multiple sources to create "diverse and enriched datasets" and that it "transmute[s] existing datasets" to generate new training data.

58. Nvidia also admits that it trains on "approximately 5.9 million audio-text pairs" curated from "highly heterogeneous data" aggregated from multiple datasets.

59. Nvidia acknowledges that its model development depends on "collecting and training on highly heterogeneous data," including the combination of numerous datasets from disparate sources.

60. Despite these admissions, Nvidia fails to demonstrate compliance with the licensing terms governing those datasets, including the MTG-Jamendo Dataset. Nvidia's publicly-available publications expressly identify MTG-Jamendo as one of the datasets Nvidia utilized for its Fugatto and Audio Flamingo products. Such use was commercial in nature, being directed toward commercial advantage or monetary compensation and furthermore done without permission or authorization from Jamendo, including through the CC BY-NC 4.0 license.

61. By aggregating, copying, transforming, and "transmuting" such datasets, Nvidia necessarily reproduced, stored, and created derivative works of copyrighted material, including the MTG-Jamendo Dataset, in violation of applicable licensing restrictions.

62. Nvidia's use of such data was undertaken for commercial purposes, including the

development and deployment of its generative AI technologies, and without obtaining the required permissions from Jamendo. Such conduct constitutes willful copyright infringement of the MTG-Jamendo Dataset and the Jamendo Database from which it is derived.

63. Upon information and belief, Nvidia copied, ingested, stored, and processed the MTG-Jamendo Dataset, including its structure, organization, and associated metadata, within its machine learning training systems.

64. Additionally, upon information and belief, Nvidia copied, ingested, stored, and processed various audio tracks listed in the Jamendo Database and the MTG-Jamendo Dataset, including many tracks for which Jamendo holds an exclusive right to commercial exploitation. By copying such musical works and sound recordings for commercial purposes, NVIDIA has infringed upon the exclusive rights under the Copyright Act of 1976 in those audio tracks. Upon information and belief, the artists of such tracks exclusively licensed to Jamendo have not granted permission to any third party to exploit such tracks commercially. As Jamendo holds the exclusive right to commercial exploitation, Jamendo has standing to assert claims of copyright infringement falling within the scope of Jamendo's exclusive license, including at least the exclusive right to reproduce, modify, fix on any media, publish, broadcast, distribute or transfer in any format the musical works and sound recordings.

65. In its pursuit of commercial success through the development of its generative AI technologies, NVIDIA has disregarded Jamendo's intellectual property rights by using the MTG-Jamendo Dataset, the Jamendo Database and various musical works and sound recordings listing within the Jamendo Database including those listed in **Exhibit A** without authorization. Such conduct constitutes infringement of Jamendo's copyrights through the unauthorized use of the database, dataset, and musical works and sound recordings in the development, training, and operation of commercial AI systems.

66. In or around March 2024, Jamendo first learned that Nvidia had utilized the Jamendo Dataset and/or MTG-Jamendo Dataset in the development of Fugatto and Audio Flamingo. Specifically, Nvidia explicitly recognized that it used Jamendo as a training source for Fugatto and Audio Flamingo.

67. Nvidia never contacted Jamendo, as expressly directed on the MTG-Jamendo Dataset GitHub page, to obtain a license for any commercial use of the MTG-Jamendo Dataset, the Jamendo Database, or any of the musical works and sound recordings listed in the same. As set forth above, the applicable Creative Commons license does not permit use of the MTG-Jamendo Dataset for commercial purposes. Nvidia's failure to obtain such authorization renders its use of the dataset unauthorized and in violation of the governing license terms.

Jamendo's Licensing Efforts and Nvidia's Refusal

68. In or around mid-2024, after learning of Nvidia's use of the MTG-Jamendo Dataset as well as the musical works and sound recordings listed in **Exhibit A** for commercial purposes, Jamendo sought to initiate discussions with Nvidia regarding a commercial license.

69. From 2024 through June 2025, Jamendo sought to resolve the matter through multiple efforts at negotiation with Nvidia and its representatives, consistent with its standard practice of licensing datasets for AI and machine learning uses.

70. On or around September 29, 2025, Jamendo issued invoices to both Nvidia Technologies Belgium and Nvidia for payment for commercial use of the MTG-Jamendo dataset. Attached hereto as **Exhibit D** is a true and correct copy of Jamendo's invoices to Nvidia Technologies Belgium and Nvidia. Subsequent reminders and notices of default were sent to both Nvidia Technologies Belgium and Nvidia. Neither Nvidia Technologies Belgium nor Nvidia challenged the invoices or the subsequent reminders upon receipt. Nvidia eventually responded, after the notices of default had been sent, by asserting that no contractual relationship had ever

existed between the parties. Attached hereto as **Exhibit E** is a true and correct copy of Nvidia's denial of any contractual relationship or obligation to pay the invoices issued by Jamendo.

71. Nvidia has achieved unprecedented financial success through the commercialization of AI technologies built on large-scale datasets, such as the MTG-Jamendo Dataset, yet it declined to compensate or even meaningfully engage with Jamendo regarding the use of its dataset, highlighting a stark disconnect between profit and accountability.

72. Nvidia's own leadership has recognized that AI constitutes "the foundational infrastructure of the modern world" and that its responsible deployment is critical to shaping the future. Yet, despite these public pronouncements, Nvidia failed to engage with Jamendo after being notified of its unauthorized use of the MTG-Jamendo Dataset.

73. Upon information and belief, NVIDIA has continued to use the MTG-Jamendo Dataset despite receiving notice that such use for commercial purposes requires a license from Jamendo.

74. Nvidia has failed to cure its default, has not paid the invoiced amount, and has continued to retain the benefits of its unauthorized use of the MTG-Jamendo Dataset.

Jamendo's Belgian Action Against NVIDIA Technologies Belgium

75. On October 14, 2025, Jamendo's counsel sent a formal notice requesting payment of approximately EUR 17.8 million, including default interest (12%) and a lump-sum indemnity (10%).

76. On November 7, 2025, NVIDIA Technologies Belgium was formally summoned to appear before the Belgian court (Ghent).

77. Jamendo's claim is based on its general terms and conditions, which are referenced in the invoice issued to NVIDIA Technologies Belgium.

78. NVIDIA Technologies Belgium has challenged both (i) the existence of any

contractual relationship with Jamendo and (ii) the international jurisdiction of the Belgian court.

79. At the introductory hearing on November 20, 2025, an objection to jurisdiction was raised, and it was agreed that this issue would be addressed as a preliminary matter before any discussion on the merits.

80. A hearing on jurisdiction took place on March 26, 2026 before the Ghent court and on June 11, 2026, the Court rejected Nvidia Belgium's jurisdictional objection, ordering that the matter shall proceed and be heard on its merits. Still, Nvidia Belgium disputes the validity and enforceability of Jamendo's invoice.

Count I

Copyright Infringement in Violation of 17 USCS § 101 *et seq*.

81. Jamendo repeats and realleges the allegations set forth in the preceding paragraphs as though fully set forth herein.

82. Jamendo owns valid and enforceable rights in a copyrighted compilation consisting of the MTG-Jamendo dataset, including its selection, coordination, and arrangement of metadata and associated content.

83. On or around June 17, 2026, the U.S. Copyright Office issued Registration No. TX-9-606-566 concerning the MTG-Jamendo Dataset. Jamendo is the owner and holder of Registration No. TX-9-606-566 issued by the US Copyright Office. Attached hereto as **Exhibit F** is a true and correct copy of the copyright registration dated June 17, 2026.[1]

84. The MTG-Jamendo Dataset embodies Jamendo's structured metadata systems, including tagging frameworks, classification methods, and organizational schema developed and maintained through substantial investment and ongoing effort. The MTG-Jamendo Dataset was created by Jamendo from the Jamendo Database and is a derivative work thereof.

[1] Jamendo has filed an application to register the Jamendo Database with the U.S. Copyright Office and will request to amend this Complaint once a registration has been issued.

85. The MTG-Jamendo Dataset was made available under the CC BY-NC 4.0 license, which permits use only for non-commercial purposes and prohibits use directed toward commercial advantage or monetary compensation.

86. Nvidia used the MTG-Jamendo Dataset to develop, train, and operate its commercial audio AI systems, Fugatto and Audio Flamingo.

87. Upon information and belief, Nvidia's use of the MTG-Jamendo Dataset constitutes commercial use within the meaning of the CC BY-NC 4.0 license and exceeds the scope of any permission granted.

88. Because Nvidia's use falls outside the scope of the CC BY-NC 4.0 license, Nvidia's copying and use of the MTG-Jamendo dataset was unauthorized.

89. Upon information and belief, Nvidia reproduced Jamendo's copyrighted MTG-Jamendo Dataset, including by copying, storing, and processing the MTG-Jamendo Dataset in whole or in substantial part within its machine learning training systems.

90. Such acts constitute reproduction and use of the MTG-Jamendo Dataset without authorization in violation of 17 U.S.C. § 106(1).

91. Nvidia's infringement has been and continues to be willful, including because Nvidia had notice of the licensing restrictions governing the MTG-Jamendo Dataset and continued its conduct notwithstanding such notice.

92. As a direct and proximate result of Nvidia's infringement, Jamendo has suffered and will continue to suffer irreparable harm, including loss of control over its copyrighted compilation, diminution in the value of the MTG-Jamendo Dataset, and loss of licensing revenue.

93. Unless enjoined by this Court, Nvidia will continue to infringe Jamendo's rights. Jamendo is therefore entitled to injunctive relief.

94. Jamendo is further entitled to recover its actual damages and any additional profits

of Nvidia attributable to the infringement, or, in the alternative, statutory damages pursuant to 17 U.S.C. § 504(c), as well as its costs and reasonable attorneys' fees pursuant to 17 U.S.C. § 505.

Count II

Copyright Infringement in Violation of 17 USCS § 101 *et seq*.

95. Jamendo repeats and realleges the allegations set forth in the preceding paragraphs as though fully set forth herein.

96. Jamendo is the holder of exclusive, worldwide, transferable license to numerous musical works and sound recordings present within Jamendo's commercial licensing business. These exclusive licenses have been issued by various artists. The exclusive licenses include the exclusive rights to reproduce, modify, fix, publish, broadcast, distribute or transfer in any format, publicly communicate, perform, or broadcast, including through digital audio means. For the majority of the works for which Jamendo holds an exclusive license, the artist has issued an underlying open license allowing specific, limited uses of the musical work at issue, such as a Creative Commons license. While such licenses permit certain uses by third parties, such uses have specific conditions and limitations. A common restriction that Jamendo's artist licensors have included in their underlying open licenses is a restriction against commercial uses of their musical works and sound recordings. Such restrictions and limitations are not included in Jamendo's exclusive licenses which expressly include the exclusive right to promote the artist or Jamendo, to be included in Jamendo's Licensing Commercial Program, and even to be used "for the purposes of text and data mining, machine learning, research, training, development, and testing of next-generation AI models, products, services, software, and tools. This includes the commercial and/or non-commercial use and launch of such AI models, products, services, software, and tools." Jamendo thus holds exclusive rights to all uses of the musical works and sound recordings at issue beyond those granted in any applicable underlying license. Jamendo's

ownership of the exclusive license to the musical works and sound recordings grants it standing to bring claims of copyright infringement for any uses within the scope of its exclusive license. *See* 17 U.S.C. § 501(b); *Minden Pictures, Inc. v. John Wiley & Sons, Inc.* (9th Cir. 2015) 795 F.3d 997, 1002.

97. Attached as **Exhibit A** is a list of musical works and sound recordings for which Jamendo holds an exclusive license as described above, and for which the artists have issued an underlying license with a restriction on commercial uses according to the underlying license identified with each musical work, and the information concerning the registration of each musical work with the U.S. Copyright Office. The information contained in **Exhibit A** is incorporated as if set forth herein.

98. Each musical work listed in **Exhibit A** embodies a musical work as described in 17 U.S.C § 102(a)(2).

99. Upon information and belief, Nvidia used some or all of the musical works and sound recordings listed in **Exhibit A** to develop, train, and operate its commercial audio AI systems, Fugatto and Audio Flamingo. Nvidia's use of the tracks at issue was intended for or directed toward commercial advantage or monetary compensation, including uses generating revenue through advertising, affiliate programs, or product integration, as well as uses that confer a commercial advantage by reducing research and development costs or accelerating the development of commercial products and services, regardless of whether a particular system has been publicly released

100. Upon information and belief, Nvidia's uses of the musical works and sound recordings listed in **Exhibit A** constitutes commercial use within the meaning of the CC BY-NC 4.0 license and exceeds the scope of any permission granted. Because Nvidia's use falls outside the scope of the CC BY-NC 4.0 license, Nvidia's copying and use of the MTG-Jamendo dataset

was unauthorized, and infringing upon Jamendo's exclusive rights.

101. Upon information and belief, Nvidia reproduced the musical works and sound recordings in **Exhibit A**, including by copying, storing, and processing the musical works and sound recordings in whole or in substantial part within its machine learning training systems.

102. Such acts constitute reproduction and use of the musical works and sound recordings without authorization in violation of 17 U.S.C. § 106(1).

103. Nvidia's infringement has been and continues to be willful, including because Nvidia had notice of the licensing restrictions governing the MTG-Jamendo Dataset and continued its conduct notwithstanding such notice.

104. As a direct and proximate result of Nvidia's infringement, Jamendo has suffered and will continue to suffer irreparable harm, including loss of control over its copyrighted musical works and sound recordings, diminution in the value of its musical works and sound recordings, and loss of licensing revenue.

105. Unless enjoined by this Court, Nvidia will continue to infringe Jamendo's rights. Jamendo is therefore entitled to injunctive relief.

106. Jamendo is further entitled to recover its actual damages and any additional profits of Nvidia attributable to the infringement, or, in the alternative, statutory damages pursuant to 17 U.S.C. § 504(c), as well as its costs and reasonable attorneys' fees pursuant to 17 U.S.C. § 505.

Count III

Breach of Contract – CC BY-NC 4.0 License

107. Jamendo repeats and realleges the allegations set forth in the preceding paragraphs as though fully set forth herein.

108. The MTG-Jamendo dataset was made publicly available subject to CC BY-NC 4.0 license, together with the accompanying notice, stating:

"Please note that the MTG-Jamendo Dataset is made available solely for non-commercial research and academic use. Any other use, including but not limited to commercial applications, requires prior written authorization from Jamendo S.A. For any commercial use requests, please contact Jamendo's licensing team at hello@jamendo.com to obtain a license adapted to your project."

109. The CC BY-NC 4.0 license grants permission to use the MTG-Jamendo Dataset subject to express conditions, including a restriction limiting use to non-commercial purposes.

110. The MTG-Jamendo Dataset was further distributed with the clear notice, through the GitHub repository, stating that it is made available solely for non-commercial research and academic use, and that any commercial use requires prior written authorization from Jamendo

111. These notices define commercial use to include any use intended for or directed toward commercial advantage or monetary compensation, including uses generating revenue through advertising, affiliate programs, or product integration, as well as uses that confer a commercial advantage by reducing research and development costs or accelerating the development of commercial products and services, regardless of whether a particular system has been publicly released.

112. Upon information and belief, by accessing, downloading, and using the MTG-Jamendo Dataset, Nvidia accepted and assented to the governing terms and conditions, including both the CC BY-NC 4.0 license and the accompanying restrictions and notices.

113. Upon information and belief, Nvidia breached the license by using the MTG-Jamendo Dataset for commercial purposes without obtaining authorization from Jamendo.

114. Upon information and belief, Nvidia's use of the MTG-Jamendo Dataset for commercial purposes violates the express non-commercial restrictions set forth in the CC BY-NC 4.0 license and the accompanying notice.

115. The governing terms and restrictions were intended to protect the commercial

value of the MTG-Jamendo Dataset and Jamendo's licensing interests.

116. Jamendo is an intended beneficiary of those terms and restrictions.

117. Upon information and belief, Nvidia's conduct constitutes a material breach of the governing terms and conditions.

118. As a direct and proximate result of Nvidia's breach, Jamendo has suffered damages, including but not limited to, lost licensing revenue and diminution in the value of the MTG-Jamendo Dataset

119. Nvidia's breach has caused and will continue to cause irreparable harm to Jamendo, for which Jamendo has no adequate remedy at law.

120. Jamendo is entitled to all available remedies for breach of contract, including damages and such further relief as the Court deems just and proper.

Count IV

Breach of Contract – Terms of Use

121. Jamendo repeats and realleges the allegations set forth in the preceding paragraphs as though fully set forth herein.

122. Jamendo allows users to make queries to the Jamendo Database and access the music tracks listed within the Jamendo Database at its website, located at www.jamendo.com (the "Jamendo Website").

123. Use of the Jamendo Website is governed by written terms and conditions identified as the Terms of Use. Such Terms of Use have been located at https://www.jamendo.com/legal/terms-of-use. The Terms of Use governing Jamendo's Website have always restricted the type and manner of use of the content found on and through the Jamendo Website.

124. As a condition of use of the Jamendo Website, all users are required to be bound

by the Terms of Use. Specifically, the Terms of Use state that:

This is a contract between you and Jamendo. By accessing, visiting, or using the Jamendo website and services, whether or not you register, you agree to be bound by these Terms of Use. If you are using the Jamendo website on behalf of your employer or a company, you confirm that you have the authority to bind your employer/company to these Terms of Use and any License/Subscription you may purchase. References to "you" and "your" in these Terms of Use will apply to both you and your employer/company, as applicable. Do not agree to these Terms or any License on behalf of your employer/company if you are not authorized to do so. Please read them carefully. If you do not accept these Terms of Use, do not use Jamendo's website or services.

125. Users must affirmatively accept and agree to be bound by the Terms of Use, in order to be permitted to access the Website content.

126. The Terms of Use at Jamendo's Website further prohibit the use of content from Jamendo's Website for data mining and machine learning, among other uses:

Users are prohibited from using any content from Jamendo for commercial text and data mining purposes, including but not limited to machine learning, artificial intelligence, or any other automated data analysis techniques. Such unnauthorized use may result in legal action. Content from Jamendo may be used for research and scientific purposes, only for non-commercial use, under certain circumstances. We request that you seek prior written authorisation from Jamendo before using Jamendo content for such purposes.

127. Upon information and belief, Nvidia accessed, used, and downloaded information and music tracks from the Jamendo Website and thereby accepted and assented to the Terms of Use listed on the Jamendo Website.

128. Upon information and belief, Nvidia breached the Terms of Use from the Jamendo Website by using the information and music tracks downloaded from the Jamendo Website for

commercial purposes without obtaining authorization from Jamendo.

129. Upon information and belief, Nvidia's use of the information and music tracks downloaded from the Jamendo Website violates the express restrictions set forth in the Terms of Use.

130. Nvidia's conduct constitutes a material breach of the governing Terms of Use of the Jamendo Website.

131. As a direct and proximate result of Nvidia's breach, Jamendo has suffered damages, in an amount to be proved at trial.

132. Nvidia's breach has caused and will continue to cause irreparable harm to Jamendo, for which Jamendo has no adequate remedy at law.

133. Jamendo is entitled to all available remedies for breach of contract, including damages and such further relief as the Court deems just and proper.

Count V

Unjust Enrichment

134. Jamendo repeats and realleges the allegations set forth in the preceding paragraphs as though fully set forth herein.

135. Jamendo has developed and maintains a valuable dataset reflecting substantial investment in the creation, organization, and maintenance of musical content, metadata systems, and classification structures.

136. The MTG-Jamendo Dataset embodies its structured organization of metadata, including tagging frameworks, classifications, and curated arrangements developed through significant time, effort, and expertise.

137. Jamendo has also developed a commercial market for licensing its catalogue, metadata, and related datasets for AI and machine learning purposes.

138. Upon information and belief, Nvidia obtained and used the MTG-Jamendo Dataset in connection with the development, training, and operation of its commercial artificial intelligence systems, Fugatto and Audio Flamingo.

139. Upon information and belief, by using the MTG-Jamendo dataset, Nvidia obtained substantial benefits, including immediate access to a large, organized, and commercially useful catalogue of rights-cleared music with detailed metadata, classifications, and tags suitable for AI training.

140. Upon information and belief, Nvidia also avoided the substantial time, cost, and effort required to build or license a comparable dataset, including acquiring rights from thousands of artists, organizing and maintaining a music database, validating metadata, and developing classification and tagging systems.

141. Upon information and belief, Nvidia has profited, and continues to profit, from its use of the MTG-Jamendo Dataset, including through the commercialization of its Audio Flamingo model and through the development of its Fugatto model, which provides commercial advantage by reducing research and development costs and accelerating Nvidia's AI offerings.

142. Nvidia's retention of these benefits without compensation to Jamendo is inequitable.

143. Jamendo reasonably expected to be compensated for commercial uses of the MTG-Jamendo Dataset, as reflected in its licensing practices and market negotiations for AI-related uses.

144. Upon information and belief, Nvidia knowingly retained these benefits without payment to Jamendo.

145. Nvidia's retention of such benefits is unjust and inequitable.

146. As a direct and proximate result of Nvidia's conduct, Jamendo has suffered harm,

including lost licensing revenue and diminished market value of its dataset.

147. Jamendo is entitled to restitution and disgorgement of the benefits unjustly retained by Nvidia.

Count VI

Unfair Competition – Cal. Bus. & Prof. Code § 17200

148. Jamendo repeats and realleges the allegations set forth in the preceding paragraphs as though fully set forth herein.

149. Nvidia has engaged in unlawful business acts and practices within the meaning of California Business and Professions Code § 17200.

150. Nvidia's conduct constitutes unlawful business practices in that it violates, inter alia, the Copyright Act, 17 U.S.C. § 101 et seq., and breaches binding licensing terms governing the MTG-Jamendo Dataset, including the CC BY-NC 4.0 license and accompanying restrictions.

151. By engaging in conduct that violates these statutory and contractual obligations, Nvidia has committed "unlawful" acts within the meaning of Section 17200.

152. Nvidia's conduct is unfair within the meaning of Section 17200.

153. Nvidia has knowingly used the MTG-Jamendo Dataset for commercial AI training without authorization, thereby circumventing Jamendo's established licensing framework for AI and machine learning uses.

154. Nvidia operates in a commercial market in which datasets such as the MTG-Jamendo Dataset are licensed for compensation based on scope, scale, and intended use.

155. Nvidia's conduct allows it to obtain the benefit of the MTG-Jamendo Dataset without incurring the costs that lawful market participants must bear.

156. Upon information and belief, by avoiding the costs of licensing or independently developing a comparable dataset, Nvidia has reduced its research and development expenses and

accelerated the development of its AI systems.

157. Nvidia's conduct thereby confers an unfair competitive advantage in the market for AI technologies, including audio generation and audio-language systems.

158. Nvidia's conduct undermines and threatens the viability of the market for licensed music datasets used in AI training.

159. Nvidia's conduct violates the policy and spirit of laws protecting fair competition and intellectual property, and is immoral, unethical, oppressive, and unscrupulous.

160. Nvidia's conduct is also fraudulent within the meaning of Section 17200.

161. Nvidia has represented, directly or by implication, that its AI systems are developed using lawfully obtained or permissible training data. In fact, upon information and belief, Nvidia has used the MTG-Jamendo Dataset subject to licensing restrictions without complying with those restrictions.

162. Such conduct is likely to deceive consumers, business partners, and market participants regarding the legality and provenance of Nvidia's training data.

163. Jamendo has suffered injury in fact and has lost money or property as a result of Nvidia's conduct. Jamendo's injuries include, but are not limited to, lost licensing revenue, lost business opportunities, and diminution in the value of the MTG-Jamendo Dataset.

164. Nvidia's conduct has directly caused these injuries by depriving Jamendo of compensation it would otherwise have received in the ordinary course of its licensing business.

165. Jamendo is entitled to restitution of all monies wrongfully obtained by Nvidia as a result of its unfair competition.

166. Jamendo is further entitled to disgorgement of Nvidia's ill-gotten gains.

167. Jamendo is entitled to injunctive relief prohibiting Nvidia from continuing its unlawful and unfair conduct.

PRAYER FOR RELIEF

WHEREFORE, Jamendo respectfully requests that this Court enter judgment in its favor and against Nvidia, and grant the following relief:

1. A judgment declaring that Nvidia has infringed Jamendo's copyrighted compilation and musical works and sound recordings in violation of 17 U.S.C. §§ 106 and 501;

2. A judgment declaring that Nvidia has breached the terms and conditions governing use of the MTG-Jamendo dataset, including the CC BY-NC 4.0 license;

3. A judgment declaring that Nvidia has been unjustly enriched at Jamendo's expense;

4. A judgment declaring that Nvidia has engaged in unlawful, unfair, and/or fraudulent business acts and practices in violation of California Business and Professions Code § 17200;

5. Preliminary and permanent injunctive relief prohibiting Nvidia, and its officers, agents, servants, employees, attorneys, and all persons acting in concert or participation with it, from directly or indirectly copying, using, reproducing, or otherwise exploiting the MTG-Jamendo dataset, or any portion thereof, as well as any of the musical works and sound recordings listed in **Exhibit A** for commercial purposes without authorization;

6. An order requiring Nvidia to cease use of the MTG-Jamendo Dataset as well as any of the musical works and sound recordings listed in **Exhibit A** in connection with its artificial intelligence systems and, to the extent required by law, to delete or destroy copies of the dataset and any materials derived therefrom;

7. An award of Jamendo's actual damages and Nvidia's profits attributable to the infringement pursuant to 17 U.S.C. § 504(b) in an amount no less than EUR 17.8 million, or, in the alternative, statutory damages pursuant to 17 U.S.C. § 504(c), which provides for statutory

damages of up to USD 30,000 per infringed work, and up to USD 150,000 per infringed work in cases of willful infringement, subject to the Court's determination and applicable law;

8. A finding that Nvidia's infringement was willful, entitling Jamendo to enhanced statutory damages;

9. An award of restitution and disgorgement of all revenues, profits, and other benefits unjustly obtained by Nvidia as a result of its conduct;

10. An award of damages arising from Nvidia's breach of contract, including but not limited to the fair market value of a commercial license for Nvidia's use of the MTG-Jamendo Dataset as well as the musical works and sound recordings listed in **Exhibit A**;

11. An order awarding Jamendo restitution and injunctive relief under California Business and Professions Code § 17200;

12. An award of Jamendo's costs and reasonable attorneys' fees pursuant to 17 U.S.C. § 505 and any other applicable law;

13. An award of pre-judgment and post-judgment interest as permitted by law;

14. Such other and further relief as the Court deems just and proper.

JURY DEMAND

Jamendo demands a trial by jury on all issues so triable.

Dated: June 23, 2026

ROPERS MAJESKI PC

By: /s/ Kevin W. Isaacson
TODD A. ROBERTS
KEVIN W. ISAACSON
Attorneys for Plaintiff
S.A. JAMENDO, a company formed under the laws of Grand Duchy of Luxembourg

Exhibit A

No.	Copyright Registration No.	Title of Work	Contents	Copyright Claimant	Registration Date	Type of Work
1	SR0000813563	Get Up And Move!	Dazed But Not Confused Get Up And Move! Juicy Sound Waves Mood SWings Never Forget (Vocal Ecliptic Mix) CUMMMBIA (Luscious P.M. Deep House Rrrmix) CUMMMBIA (Luscious P.M. Deep House Rrrmix) (Extended Mix) CUMMMBIA CUMMMBIA (Extended Mix) Faq Faq (Hip-Hop Instrumental Mix) Coming To You Straight Coming To You Straight (Instrumental) Coming To You Straight (Hip Hop Instrumental Mix) Question Mark In the Name of Fun If It Wasn't For This Bass (Ibiza A.M. Bounce Mix) Lazer Lightz Project541 EDM (BEAT to the BOX mix) Page 2 of 3 Stereo Delight Glimmering Lights We're Down To Get Down (Deep House On Fire Remix) Esta es mi ciudad (C'est ma ville This Is My City) Esta es mi ciudad (C'est ma ville This Is My City)(Extended Mix)	Michael Martinez	July 31, 2017	Sound Recording and Music

No.	Copyright Registration No.	Title of Work	Contents	Copyright Claimant	Registration Date	Type of Work
			Breakdancing To My Heartbeat Arcu Iris			
No.	**Copyright Registration No.**	**Title of Work**	**Contents**	**Copyright Claimant**	**Registration Date**	**Type of Work**
2	SR0000754677	Rainbow Stone, et al.	Rainbow Stone. Kids in Toyland. Surfing On Gasoline. Burning Sky. Turn The Lights On. Fireballs. Wizard. One Step Away. Warm Up. Mad Jungle. Endless Hours. Rainbow Stone (Remix) Moonlight Dancer. What's The Matter With You. Stainless Steel Machine.	Pierre Rousseau	December 13, 2014	Sound Recording and Music
No.	**Copyright Registration No.**	**Title of Work**	**Contents**	**Copyright Claimant**	**Registration Date**	**Type of Work**
3	SR0000779257	White Silk Toga	White Silk Toga. My Own Fantasy. Superman Kid. Catch the Sphere. Puzzle. Twister Illusion. Blazing Sunlight. Happy Monkeys. Crystallized. What's the Combination. Last Day. Ruins. Stainless Steel Machine. Devil's Playground.	Pedro Rousseau, pseud. of Pierre Rousseau, (author of pseudonymous work)	February 6, 2016	Sound Recording

No.	Copyright Registration No.	Title of Work	Contents	Copyright Claimant	Registration Date	Type of Work
			Ritournelle.			
No.	**Copyright Registration No.**	**Title of Work**	**Contents**	**Copyright Claimant**	**Registration Date**	**Type of Work**
4	SRU001138121	Apogalactica	Apogalactica. Cosmic Storm. Funky Funk. Jambalaya. Lovely Monster. Misty Rain. Never Say Never. Super Nova. The Quest. Two Hearts One Life	Abel Quintero	September 23, 2013	Sound recording and music
No.	**Copyright Registration No.**	**Title of Work**	**Contents**	**Copyright Claimant**	**Registration Date**	**Type of Work**
5	SRU000648493	Make this end		© ℗ on words, music, production, and performance; Pierre Rousseau-	January 31, 2006	Sound recording and music

Exhibit B

Fugatto 1

Foundational Generative Audio Transformer Opus 1

NVIDIA
Rafael Valle, Rohan Badlani, Zhifeng Kong, Sang-gil Lee, Arushi Goel, Sungwon Kim, João Felipe Santos, Shuqi Dai, Siddharth Gururani, Aya AlJa'fari, Alex Liu, Kevin Shih, Wei Ping, Bryan Catanzaro
rafaelvalle@nvidia.com

ABSTRACT

Fugatto is a versatile audio synthesis and transformation model capable of following free-form text instructions with optional audio inputs. While large language models (LLMs) trained with text on a simple next-token prediction objective can learn to infer instructions directly from the data, models trained solely on audio data lack this capacity. This is because audio data does not inherently contain the instructions that were used to generate it. To overcome this challenge, we introduce a specialized dataset generation approach optimized for producing a wide range of audio generation and transformation tasks, ensuring the data reveals meaningful relationships between audio and language. Another challenge lies in achieving compositional abilities – such as combining, interpolating between, or negating instructions – using data alone. To address it, we propose *ComposableART*, an inference-time technique that extends classifier-free guidance to compositional guidance. It enables the seamless and flexible composition of instructions, leading to highly customizable audio outputs outside the training distribution. Our evaluations across a diverse set of tasks demonstrate that *Fugatto* performs competitively with specialized models, while *ComposableART* enhances its sonic palette and control over synthesis. Most notably, we highlight our framework's ability to synthesize emergent sounds – sonic phenomena that transcend conventional audio generation – unlocking new creative possibilities. Demo Website.

1 INTRODUCTION

Recent research has sparked a strong debate between **specialist** and **generalist** models. While specialist models excel at specific tasks, they tend to be brittle, often struggling with changes in data distribution or task requirements. In contrast, generalist models eliminate the need for task-specific designs, can process diverse data, and scale effectively with increased compute and data. They also demonstrate emergent capabilities, enabling unsupervised task learning by leveraging broader datasets (Radford et al., 2019) and demonstrations (Brown et al., 2020). In this paper, we propose a strategy for developing a generalist audio synthesis and transformation model, called *Fugatto*, and an inference method for composing instructions through latent space manipulation, including from different models, called Composable Audio Representation Transformation (*ComposableART*).

Large language models (LLMs) have demonstrated impressive unsupervised multitask learning capabilities in the text domain (Radford et al., 2019), where instructions can be inferred from the data itself. However, such instructions are typically absent in the audio domain, making it difficult to generalize to unseen tasks without explicit guidance. Although models such as (Wang et al., 2024; Yang et al., 2023; Vyas et al., 2023) exist, they have several limitations enumerated in Table 1. In this panorama, dataset and instruction generation is necessary.

We employ a **multifaceted data and instruction generation strategy** that considerably expands the range of tasks of audio generation models. First, we use LLMs to generate and augment instructions and captions, providing *Fugatto* with instructions that are closer to free-form instructions (Goel et al., 2024; Doh et al., 2023). Second, we develop instructions that can be either absolute (e.g.,

"synthesize a happy voice") or relative (e.g., "increase the happiness of this voice"), enabling *Fugatto* to handle a wide array of dynamic tasks (OpenAI, 2024). Third, we leverage audio understanding models (Kong et al., 2024a; Gong et al., 2023) to create descriptions and synthetic captions for audio clips, enriching the dataset where annotations are sparse, allowing for better generalization and more accurate performance (Kong et al., 2024b). Fourth, we transmute existing datasets to uncover new relationships, enabling the creation of entirely new tasks without requiring additional raw data. Finally, we use audio processing tools to create new connections between text, audio, and their corresponding transformations.

By combining these approaches, we ensure that *Fugatto* has access to diverse and enriched datasets, allowing it to learn from various audio domains and contexts. This strategy enhances the model's task diversity and lays the groundwork for unsupervised multitask learning at scale, unlocking emergent abilities such as synthesizing entirely new sounds, such as "a saxophone barking".

While data is important, the ability to **compose, interpolate between, and negate instructions** is generally difficult to obtain through data alone. Negation data, for instance, is typically unavailable, and generating outputs that represent the composition or interpolation of instructions is equally challenging. Though this has been explored in the vision domain using Energy Based Models(EBMs) (Du et al., 2020) and EBMs in latent space (Nie et al., 2021), EBMs require training a new model per attribute, which can be cumbersome and impractical for a large number of attributes.

To overcome this limitation, we propose an inference-time technique called *ComposableART*, which is based on classifier-free guidance (CFG) (Ho & Salimans, 2022) and further expands on the dual classifier-free guidance in (Yang et al., 2024; Lee et al., 2024). We propose a generalized framework that leverages the weighted combination of vector fields between instructions, frame indices and models. This approach enables *Fugatto* to handle complex instruction-based operations, such as smoothly interpolating between instructions or negating specific instructions to exclude undesired effects. In contrast, models like (Vyas et al., 2023; Yang et al., 2023) rely on more rigid methods, often requiring external classifiers or manual intervention to achieve similar results.

In this paper, we present a detailed exploration of *Fugatto*'s dataset and instruction generation, training strategies, implementation improvements, and performance across a wide range of tasks. Through rigorous evaluations and comparisons with specialist and generalist models, we establish *Fugatto* as a new benchmark for foundation models for audio synthesis and transformation. Similarly, we establish *ComposableART* as a highly desirable framework for compositional guidance that unlocks the full potential of score-based generative models. Our contributions include:

- We offer a comprehensive strategy for building a foundation model for audio generation and transformation given text and audio inputs, delivering strong performance across a wide range of tasks and providing a robust framework for both research and practical applications.
- We demonstrate how to enhance and create contextually rich audio and text datasets while generating flexible instructions with LLMs, enabling our community to replicate and adapt these techniques for their own models.
- We demonstrate how to perform composition, interpolation, and negation of instructions by extending CFG to compositional guidance, enabling better control over the model's outputs.

Table 1: Comparison of our proposed model *Fugatto* with other models.

	Fugatto	AudioBox	NExT-GPT	UniAudio	AUDIT	VoiceLDM
Emergent properties	✓	?	?	?	?	?
Large-scale data	✓	✗	✓	✓	✗	✗
Supports numerous tasks	✓	✗	✗	✓	✗	✗
Free-Form Instructions	✓	✗	?	?	✓	✗
Open-ended generation	✓	✗	✗	✓	✗	✗
Compositionality	✓	✗	✗	✗	✗	✓
Multi-Modal Inputs	✓	✓	✓	✓	✓	✗

2 APPROACH

Our approach to audio generation given text and optional audio is similar to recent approaches in LLMs, focusing on large scale compute and datasets, followed by pre-training and fine-tuning stages. However, our approach differs in two aspects. First, our dataset generation mechanism requires going beyond unsupervised next token prediciton (Section 2.1). Second, we propose inference time techniques to control audio generation (Section 2.4). In Appendix A.1 we provide a detailed list of tasks and instructions our model supports.

2.1 DATASET GENERATION

We focus on building as large and diverse a dataset in order to collect demonstrations of as many audio tasks and contexts as possible. We emphasize that our ultimate goal is not to just excel on such tasks, but to drive, as a community, towards a future where unsupervised multitask learning emerges from data and model scale. Towards this goal, we propose a dataset generation strategy built on five pillars below, and provide, in Appendix A.1.1, a thorough description of each pillar.

I – Generating Free-Form Instructions with LLMs: Our strategy consists of prompting an LLM to create programmatic task-specific instruction generators, similar to Kocielnik et al. (2023). We prompt an LLM to create python methods that generate instructions of different lengths and for different personas (standard, young-crowd, thirty-somethings, professional), given task-specific inputs including audio description, language, and others. Each persona has a set of verbs, adverbs, connectors, and other assets to create the instructions.

II – Generating Absolute and Relative Instructions: Following GPT4-o's (OpenAI, 2024) ability to perform a relative change in speech, we aim to obtain an audio generation model that is able to follow instructions that are absolute or relative, such as "synthesize a happy voice" or "increase the happiness in this voice.". Given that such data is normally not available, we later describe a strategy to create such data by transmuting existing datasets and leveraging audio processing tools. Once such datasets are created, we can select one sample and create an absolute instruction for it, or select two samples and create a relative instruction. Similarly to absolute instructions, we generate relative instructions with a python method, produced by an LLM, that creates an instruction given the task, the attribute being modified and how it should be modified.

III – Generating Audio Descriptions with Audio Understanding Models: We use audio understanding and classification models (Kong et al., 2024a) to produce synthetic captions for audio in the wild, following recent research (Leng et al., 2023; Kong et al., 2024b) showing that it is possible to drastically expand or improve text-to-audio models with synthetic captions. As such, we expand the strategies described in (Leng et al., 2023; Kong et al., 2024b) to generate high quality synthetic captions. For speech data, we implemented a prompt generation pipeline that automates the creation of natural language descriptions for voices. The pipeline converts speech attributes predicted by models – such as "gender", emotion, and speech quality – into detailed natural language descriptions using LLM-generated templates. These templates describe voices in various ways based on the speaker attributes, enhancing diversity by generating descriptions in multiple formats.

IV – Creating New Tasks and Datasets by Transmuting Datasets: We leverage implicit relationships between samples in a dataset to enable new tasks. Generally speaking, we look for datasets where one factor is held constant while other factors change. For example, we leverage emotional speech synthesis datasets with different renditions of the same text (Livingstone & Russo, 2018) by the same speaker to define a speech transformation task. Similarly, we leverage instrument synthesis datasets with different renditions of the same note (Engel et al., 2017) to define an instrument transformation task. We also leverage datasets that provide the individual parts of a sounds mixture (Rafii et al., 2017) to support tasks such as source separation, and audio generation conditioned on audio context and captions, possibly synthetic.

V – Creating New Tasks and Datasets by Leveraging Audio Processing Tools: We create synthetic paired data for speech and audio by using Praat (Boersma & Van Heuven, 2001) and Pedalboard (Spotify, 2024) to manipulate several speech and audio factors. For each factor, we apply controlled modifications, allowing us to generate speech and audio samples with specific alterations. With this strategy we can create speech and audio pairs that describe speech and audio transformation tasks such as changing of one or multiple speech factors, for example "increase the F0 variance

slightly and decrease the speech rate." or "add moderate reverb to this audio file". For each factor, we determine a practical range of adjustments and define increments that correspond to varying degrees of change, such as "slightly", "moderate", and "significant".

With these pillars established and leveraging open source datasets, we are able to build a large text and audio dataset with at least 20 million rows, not including on-the-fly modifications to captions, instructions and audio. Assuming each row refers to 10 seconds of audio, our dataset is comprised of at least 50,000 hours of audio. We emphasize that this compilation is exclusively comprised of open source data, and provide a full list of datasets, tasks, and instructions in Appendices A.1.2, A.1.3 and A.1.4 respectively.

2.2 Instruction Generation

Our approach supports template-based and free-form instructions.

Template-based Instructions: In these instructions, the task is explicitly provided, followed by task-specific attributes, always in the same order, and with each attribute wrapped between start and end of attribute markers. We dynamically construct template-based instructions based on the task and the set of factors. Each factor is specified by its name and corresponding value, using the format `given {name}:{value}` followed by a closing HTML-like tag `</{factor}>`. The instruction always starts with the task, followed by its factors, and ends with `output:`. Additionally, a `given context` clause is appended at different locations, determined by the task at hand, when audio contexts are present. The structure of a template-based instruction is:

```
input:{task} given {factor}:{value}</{factor}>given context:<audio>output:
```

where {`task`} represents the specific task, {`factor`} and {`value`} correspond to the different factors and their respective values, and `<audio>` refers to the audio context provided. We provide examples of task and dataset specific instructions in Appendix A.1.4.

Free-form Instructions: We dynamically construct free-form instructions by using instruction generators introduced in Section 2.1. Unlike template-based instructions where we know before hand the reference in text to each audio context, in free-form it is not straight forward to determine which word in the text refers to the audio. As such, in free-form instructions we simply append to the instruction `given context_k:<audio>` for each audio context, resulting in this structure:

```
input:{instruction} [given context:<audio_k>]output:
```

In order to promote simplicity and agility during development, we decided to use raw text instead of learnable tokens for `given factor` and `</factor>`. Following LLM practice, the full instruction and each audio are wrapped with learnable `<start of>` and `<end of>` tokens.

2.3 Model and Training

In this section we describe the text and audio representations used in our model, as well as the training objective and architecture. We provide a graphical depiction and details for hyperparameters, objective function, training stages and oversampling in A.2.

Text and Audio representation: The text representation is obtained by encoding the previously described instructions with a pre-trained language model held frozen during training. In this *Opus*, we use the *byT5* tokenizer free language model (Xue et al., 2022), which supports a large set of characters, including IPA. In this opus, the audio representation is a 22khz mel-spectrogram with 80 bins, which is subsumed by a relatively shallow learnable transformer encoder. The mel spectrogram is scaled to have approximately 0 mean and 1 standard deviation.

Training Objective and Architecture: We train our model with the Optimal Transport Conditional Flow Matching (OT-CFM) objective (Lipman et al., 2022; Tong et al., 2023), and use a T5-based (Raffel et al., 2020) Transformer (Vaswani, 2017) with Adaptive Layer Norm (Xu et al., 2019) as the parameterization of the vector field estimator. We replace the Transformer MLPs with kernel size 3 convolutions. After independently projecting the encoded text and audio to a shared embedding space, we time-wise concatenate the embedded audio and text. The model cross-attends to this representation, applying Adaptive Layer Norm (Xu et al., 2019) to them on every layer.

We observed that certain implementation choices yielded better training curves. Specifically, adaptive layer norm is completely computed in FP32, GELU uses approximate *tanh*, and the final layers are initialized to outputs zeros, which is approximately the mean of our scaled mel-distribution.

Training Stages: *Fugatto* training follows curriculum learning[1] (Bengio et al., 2009). We start with template-based instructions and a subset of tasks. Eventually, once we informally establish that the model is able to follow template-based instructions by observing validation scores and listening to samples, we proceed with an equal mixture of template-based and free-form instructions. Given that some tasks are underrepresented in the data, we find that oversampling leads to better results. Empirically, we find that sampling from a multinomial distribution with upsampling parameter $\beta = 0.25$, similar to Le et al., 2024, is sufficient. As training progresses, we adjust each dataset's weight according to validation scores on target tasks.

2.4 Composable Audio Representation Transformation (*ComposableART*)

We extend Classifier Free Guidance(CFG) (Ho & Salimans, 2022) to support compositional guidance. Compositional guidance provides an OT-CFM model with the ability to independently control and generate (unseen) combinations of instructions and tasks, including with vector fields from different models (Karras et al., 2024). Compositional generation has been explored in Diffusion Models (Liu et al., 2023; Yang et al., 2024; Lee et al., 2024), for images and audio. To the best of our knowledge, we are the first to showcase novel ways of applying compositional guidance, expanding it to not just attributes, but also tasks, models and temporal composition of attributes.

Compositional Guidance Method Classifier Free Guidance(CFG), generally applied to diffusion models, combines the conditional and unconditional score estimates to obtain samples of higher quality and diversity pertaining to the condition. The following equation summarizes the application of CFG, where ϵ_θ represents the score estimate and γ represents the gradient scaling factor:

$$\epsilon_\theta(\mathbf{z}_\lambda, \mathbf{c}) = \epsilon_\theta(\mathbf{z}_\lambda, \mathbf{c}) + \gamma(\epsilon_\theta(\mathbf{z}_\lambda, \mathbf{c}) - \epsilon_\theta(\mathbf{z}_\lambda)) \tag{1}$$

We extend the Classifier Free Guidance framework to support the combination of vector fields across multiple instructions (c_k), multiple mel-frame indices (f) and multiple models (θ_m). Let $v_\mathrm{t,f}(c_\mathrm{k}, \theta_\mathrm{m})$ be the vector field produced at flow-step t by model m, parameterized by θ_m, given condition c_k or $\varnothing$, for mel-frame f. Additionally, let $w_\mathrm{k,f,m}$ be the flow-step invariant and user-determined scalar weight associated with $v_\mathrm{t,f}(c_\mathrm{k}, \theta_\mathrm{m})$. The equation for compositional guidance across instructions, frames, and models is defined as:

$$\tilde{v}_\mathrm{t,f} = \sum_{\mathrm{k,m}} w_\mathrm{k,f,m}(v_\mathrm{t,f}(c_\mathrm{k}, \theta_\mathrm{m}) + \gamma(v_\mathrm{t,f}(c_\mathrm{k}, \theta_\mathrm{m}) - v_\mathrm{t,f}(\varnothing, \theta_\mathrm{m}))) \tag{2}$$

where γ refers to the gradient scale parameter from CFG, and $\tilde{v}_\mathrm{t,f}$ is the resulting composed vector field for flow step t and frame f. This follows similar conditional independence assumptions in (Nie et al., 2021; Liu et al., 2023) to support compositional guidance. We apply this compositional guidance at every step of the flow-matching inference procedure.

Attribute/Event Composition: An attribute is a simple input prompt belonging to a particular task like speech synthesis. A task like audio event generation can have multiple prompts or attributes as input. With compositional guidance, we can support unique unseen combinations of attributes. This gives the users an ability to create artistic combinations like simulating a scene with multiple-audio events, e.g. by composing 'thunder', 'rain' and 'wind' a storm can be achieved.

Task Composition: The model has been trained on many individual tasks, but it has not encountered combinations of tasks, such as speech synthesis alongside audio event generation. Using compositional guidance, we can enable the synthesis of unique, unseen task combinations, like generating speech with a specific audio event in the background.

Model Composition: The same technique can be extended to integrate distinct models. This is particularly useful when training domain-specific versions of *Fugatto*, each with its own parameters

[1] It simplifies development and supports incremental research, though not strictly necessary.

and datasets. This allows users to synthesize a "mixture of experts" sample that combines models trained on independent domains such as speech and audio events. Following (Karras et al., 2024), we use the velocities for each independent model, defined by parameters θ_{m}.

Temporal Composition: Instead of using the same scalar for every frame f, we assign a unique weight $w_{\text{k,f,m}}$ to each frame. This enables users to control the compositional output with arbitrary curves (e.g., sigmoid or linear increase or decrease), while retaining the advantages of combining attributes, tasks, and models.

3 EXPERIMENTS

We present a comprehensive evaluation of *Fugatto* across multiple tasks to demonstrate its effectiveness and versatility. We begin with an ablation study, examining the impact of various design choices. Next, we evaluate *Fugatto*'s performance in audio synthesis and transformation tasks in speech, music and general sounds. Finally, we explore *Fugatto*'s emergent capabilities, and perform a thorough evaluation of our *ComposableART* method. Unless otherwise specified, we use template-based instructions, 2nd order Heun solver with 50 function evaluations, and task specific gradient scale γ.

3.1 ABLATIONS

We first analyze the impact of different t-sampling strategies in OT-CFM, comparing the traditional uniform sampling with others. Then, we examine the effect of model size on both loss metrics and emergent capabilities, including the ability to synthesize novel sounds not found in the training data—such as a "saxophone meowing" or "a person speaking while barking."

t-sampling strategy: In the OT-CFM framework, the timestep t is typically sampled from a uniform distribution, $t \sim \mathcal{U}(0, 1)$. However, recent discussions within the community have introduced conflicting strategies for this sampling process. Notably, Stable Audio's GitHub repository proposes sampling t from a sigmoid-transformed normal distribution, $t \sim \text{sigmoid}(\mathcal{N}(0, 1))$, thereby concentrating samples around $t = 0.5$. On the other hand, (Lovelace et al., 2023) advocate for increased sampling from values of t closer to zero.

In our experiments, we observe that although Stable Audio's strategy provides a marginal improvement on TTA tasks, it renders the model unable to effectively attend to the transcript in text-to-speech (TTS) tasks – a critical requirement. We provide an explanation for this phenomenon in Appendix A.3.

Training with $t \sim \mathcal{U}(0, 1)$ is effective across all tasks.
Training with $t \sim \text{sigmoid}(\mathcal{N}(0, 1))$ significantly degrades TTS performance.

Model capacity: We evaluate how the learnable parameter count influences loss curves and emergent capabilities. We consider models with 0.8 B, 1.4 B params, and 2.5 B parameters. Under a fixed data composition and sampling weights, we observe that increasing the parameter count from the smallest to the largest model not only improves validation losses but also delays overfitting. In Appendix A.2, we provide task-specific validation loss plots that showcase the expected decrease in validation loss as parameter count increases. Informally and consistent with findings in (Radford et al., 2019), we observe that the smaller model does not exhibit emergent abilities comparable to the larger models, particularly in their ability to synthesize novel sounds absent from the training data, such as "saxophone barking". We invite readers to evaluate samples in our supplementary materials.

Emerging capabilities surface with sufficient model capacity and training data.

3.2 AUDIO SYNTHESIS

Text-To-Voice Synthesis: We evaluate in-context text-to-speech synthesis (TTS) and singing voice synthesis (SVS). For TTS, we follow the evaluation in Wang et al. (2023), using the same transcripts as Eskimez et al. (2024), to evaluate our model's ability to perform speech synthesis given a transcript and a speech sample from an unseen speaker. Following our training strategy, during evaluation we always provide the speaker's previous sentence when possible, otherwise a random sample different from the target. For SVS, we evaluate *Fugatto*'s ability to generate singing voice samples from instructions describing the desired lyrics and musical style without a backing track, for example:

"Showcases a female singer with an interesting sound, conveys the message through american english lyrics, and infuses country influences throughout." The full list is available in Appendix A.4

Our zero-shot TTS results in Table 2a show that *Fugatto* has word error rates similar to ground truth, and is competitive with expert and generalist (Omni) models in terms of speaker similarity. Our SVS results in Table 2b shows high cosine similarity between CLAP embeddings of the synthetic samples and the captions used to create it. We note that the higher WER in SVS likely comes from the higher difficulty for the generative model and the speech transcription model.

Table 2: *Fugatto* is comparable to generalist models and expert models on in-context TTS benchmarks. On SVS, it synthesizes samples with high CLAP-similarity and low WER relative to the task at hand.

(a) TTS on the LibriSpeech Test Clean benchmark in Wang et al. (2023)

Model	Omni	WER ↓	SIM-o ↑	SIM-r ↑
Ground Truth		2.20		
Vall-E (24khz)	✗	5.90		0.58
Natural Speech 3 (16khz)	✗	1.81	0.67	0.76
AudioBox (16khz)	✗	4.80	0.73	
UniAudio (16khz)	✓	2.00		0.71
Fugatto $\gamma = 2$	✓	2.66	0.60	0.61
Fugatto $\gamma = 3$	✓	2.44	0.61	0.62

(b) SVS: WER and CLAP-Scores on a set of 10 music styles and 13 lyrics snippets from famous songs.

Model	WER ↓	CLAP ↑
Fugatto $\gamma = 2$	71.90	0.49
Fugatto $\gamma = 4$	19.54	0.45

Text-To-Audio (TTA): We showcase *Fugatto*'s performance on traditional TTA benchmarks that measure a model's ability to synthesize general sounds (AudioCAPS) and music (MusicCAPS) that follow instructions provided in text. We use the metrics (FD, FAD, and IS) and data splits (train, test) used in Kong et al. (2024b). Results in Table 3a and Table 3b shows that our model achieves strictly better scores than existing generalist models, while occasionally outperforming expert models.

Table 3: *Fugatto* outperforms generalists models and occasionally outperforms specialist models in TTA benchmarks on AudioCaps and MusicCaps.

(a) TTA on the AudioCaps benchmark in Kong et al., 2024b

Model	Omni	FD ↓	FAD ↓	IS ↑
VoiceLDM-Maudio	✗		2.50	
AudioBox (16khz)	✗	10.14	1.10	11.90
NExT-GPT	✓		1.68	
UniAudio	✓		3.12	
Fugatto $\gamma = 1$	✓	16.73	1.36	9.72
Fugatto $\gamma = 2$	✓	20.20	2.21	10.21

(b) TTA on the MusicCaps benchmark in Kong et al., 2024b

Model	Omni	FD ↓	FAD ↓	IS ↑
MusicGen (medium)	✗	35.52	5.02	1.94
AudioLDM-2-large	✗	16.12	2.74	2.30
Tango-AF&AC-FT-MC	✗	21.84	1.99	2.21
UniAudio*	✓		3.65	
Fugatto $\gamma = 1$	✓	11.52	1.43	2.73
Fugatto $\gamma = 2$	✓	13.18	1.93	2.97

3.3 Audio Transformations

Speech Enhancement: We evaluate speech denoising and bandwidth extension tasks. Speech denoising evaluates the ability to extract speech from an additive mixture comprised of speech and noise. Bandwidth extension (sometimes referred to as "upsampling") evaluates the ability to recreate missing content from audio that is low passed filtered and downsampled not to include frequencies above a certain threshold frequency[2]. For speech denoising, we use the DNS-Noisy benchmark and traditional metrics PESQ and STOI described in Kong et al., 2023. For bandwidth extension, we use the VCTK benchmark and the traditional metric LSD described in Liu et al., 2024. In *Fugatto*, this task is interpreted as source separation, in contrast with the enhancement task that modifies the acoustic qualities of the target audio. Though *Fugatto* is comparable to specialist models in bandwidth extension, work remains to be done to close the gap in denoising.

Speech Modulation: In this task we evaluate our model's ability to transform a person's emotion in speech into another emotion, while preserving their speaker identity and the transcript. For this purpose, construct a train and test set based on the ESD dataset. We use open-source models to report emotion classification and correlation with Valence, Arousal and Dominance (VAD). We establish

[2]We use librosa after observing that torch.audio leaks frequencies above the threshold frequency.

Table 4: *Fugatto* is comparable to specialist models for speech denoising and upsampling.

(a) Speech Denoising on the DNS benchmark in Kong et al., 2023

Model	Omni	PESQ_{WB} ↑	PESQ_{NB} ↑	STOI ↑
Noisy dataset		1.59	2.16	91.60
FullSubNet	✗	2.90	3.37	96.40
FAIR-Denoiser	✗	2.66	3.23	96.60
CleanUNet 2	✗	3.26	3.66	97.70
Fugatto $\gamma = 0.1$	✓	2.77	3.32	95.70
Fugatto $\gamma = 1$	✓	2.73	3.34	95.90

(b) Upsampling on the VCTK benchmark in Liu et al., 2024

Model	Omni	LSD 4khz ↓	LSD 8khz ↓
Unprocessed (to 22khz)		2.74	1.84
NuWave (to 22khz)	✗	1.37	0.88
NVSR (to 22khz)	✗	1.49	1.37
AudioSR (to 22khz)	✗	1.25	1.08
Fugatto $\gamma = 0.1$	✓	1.29	1.25
Fugatto $\gamma = 1$	✓	1.38	1.34

upper bounds by also computing scores on ground truth data. Our results in Table 5a show that our model is able to properly transform the emotion as well as the ground truth, it needs improvement on speaker similarity and word error rates.

MIDI2Audio: We evaluate our model's ability to convert midi to audio with the 250 simple 2-bar monophonic melodies from Pati et al., 2020. Note that *Fugatto* has never seen monophonic melodies during training, with the average number of stems present in training being 8. We use the error in pitch estimation on the ground truth MIDI notes as the upper-bound quality. We evaluate the L1 distance of the F0 contours extracted from the input rendered MIDI and the generated audio. During evaluation, we transpose the pitch contours to a common key. We provide examples in Appendix A.5.

Table 5: *Fugatto* performs well on novel tasks such as emotion conversion and MIDI2Aaudio.

(a) Speech modulation (Emotion Conversion): remarkably high Top-2 accuracy and pearson correlation ρ between ground truth and synthetic samples on VAD. Low speaker similarity.

Model	WER	SIM-o	ρ_{VAD}	Top-2 Acc.
Ground Truth	7.42			0.62
Fugatto $\gamma = 2$	24.17	0.19	0.68, 0.77, 0.77	0.59
Fugatto $\gamma = 3$	21.96	0.21	0.68, 0.77, 0.77	0.62

(b) MIDI2Audio: surprising zero-shot abilities on monophonic melody to audio.

Model	L1 ↓
F0 estimation error	0.21
Fugatto $\gamma = 1$	1.74
Fugatto $\gamma = 2$	1.00

3.4 Emergent Capabilities and Sound Gallery

We consider an ability to be emergent if it is absent in smaller models but appears in larger ones (Wei et al., 2022; Radford et al., 2019), and *Fugatto* demonstrates *emergent sounds* and *emergent tasks*. In this section, we provide qualitative results through our demo page to highlight the model's emergent abilities and invite readers to a guided tour through our sound gallery, which showcases compelling examples of *Fugatto*'s artistic potential and emergent capabilities.

Emergent Sounds: *Fugatto* exhibits the ability to generate outputs that are not present in the training data itself. For instance, it can synthesize a cello that shouts with anger or a person that speaks and barks.

Emergent Tasks: Beyond generating novel sounds, *Fugatto* demonstrates the ability to perform tasks it was not explicitly trained for by combining tasks seen during training. For example, it can perform speech prompt conditioned singing voice synthesis or convert monophonic MIDI to a singing voice.

3.5 Compositionality

Compositionality enables users to combine attributes and tasks to generate novel input combinations not found in the training data, providing artistic and fine-grained control over the desired output. Since such novel combination rarely exists in the training data or the natural world, evaluating these samples is typically challenging. Furthermore, there are no established baselines and metrics for such sounds. Despite these challenges, we aim to provide a qualitative and quantitative evaluation of our proposed approach and invite readers to listen to compositional samples on our website.

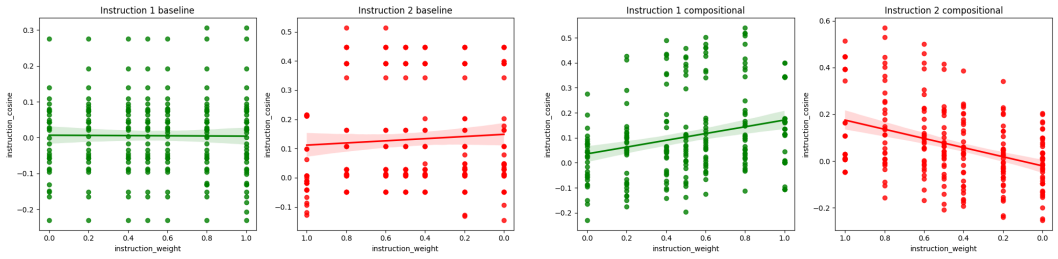
3.5.1 ATTRIBUTE/EVENT COMPOSITION

Control intensity of each instruction (Weighted Combination): Compositional synthesis with instructions gives users a knob to control the intensity of each instruction. In order to evaluate this ability, we create $\binom{10}{2}$ pairs of instructions by leveraging 10 event labels provided in Appendix A.6. Given a pair of events we can generate composite instructions through language or *ComposableART*:

Baseline linguistic composition example input:
`input: synthesize Event 1 and Event 2`
Proposed *ComposableART* composition example inputs:
`w1*v(input: synthesize Event 1) + w2*v(input: synthesize Event 2)`

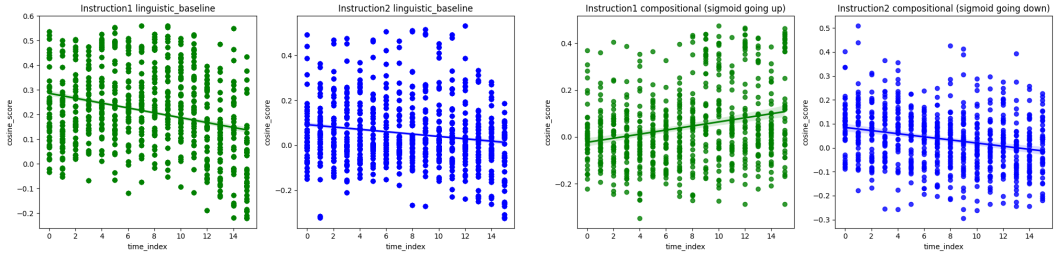
For each pair of events, we first generate samples using *ComposableART* with different convex combination of weights for each instruction, and using the linguistic baseline[3]. Then, for each method and generated audio, we compute the cosine similarity between the audio and text CLAP embeddings for each event in the event pair used to produce the audio, shown as Instruction 1 and 2 in Figure 1.

Figure 1 shows that as we increase the weight on Instruction 1, the occurrence of the event in the generated clip increases as evidenced by the CLAP scores. Reciprocally, as we decrease the weight on Instruction 2, the occurrence of the event in the generated clip decreases. In the linguistic baseline, we cannot control the weights of each attribute and their independent existence in synthesized samples is lower than the *ComposableART* samples at higher weights.



Linguistic baseline *ComposableART* (proposed).

Figure 1: Comparison of CLAP scores between the Linguistic baseline and *ComposableART*'s composition of attributes with *Fugatto*. Instruction is equivalent to event.



Linguistic Baseline *ComposableART* (proposed)

Figure 2: Comparison of Baseline and *ComposableART* temporal guidance for instruction sequences.

Negation of attributes: With our approach, users can assign negative weights to attributes, producing negative velocity that steers the model away from the attribute. We use the previous $\binom{10}{2}$ event pairs to evaluate our approach alongside linguistic negation with the keyword 'NOT':

Baseline Linguistic Negation example input:
`input: synthesize Event 1 and NOT Event 2`

[3]In Figure 1, the diagonal fit, different weights, and the lack of samples in the linguistic baseline are an implementation and compute timeout byproduct given that the linguistic baseline does not support weights.

ComposableART Negation example input:
`v(input: synthesize Event 1) - v(input: synthesize Event 2)`

Table 6 shows CLAP cosine similarity of the linguistic baseline against the proposed *ComposableART* method. It can be observed that while the positive event remains similar to the baseline, the negative event has a considerably lower cosine similarity than the baseline, indicating the effectiveness in removal of the audio event using *ComposableART* as compared to the linguistic approach. Qualitatively, we also observe that this can be immensely useful in steering towards negative emotions in speech synthesis or swapping gender, a task which is not trivial.

Interpolation of attributes: *ComposableART* also supports interpolation of attributes by having the ability to independently control one attribute while keeping others. We evaluate our ability to control "pitch" as the interpolation variable by changing the weights on pitch and keeping "text" and "language" attributes fixed. Figure 3 showcases the expected decrease in fundamental frequency as we increase the weight on the "low pitch" attribute.

Table 6: CLAP Cosine Similarity of Attributes

Instruction Type	**CLAP Score**
Linguistic baseline +ve Instruction	0.02 ± 0.02
Linguistic baseline −ve Instruction	0.17 ± 0.02
ComposableART +ve Instruction	0.06 ± 0.01
ComposableART −ve Instruction	-0.04 ± 0.02

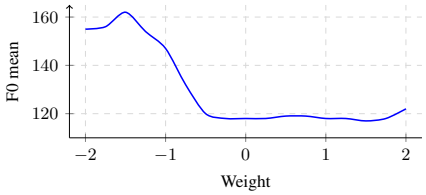


Figure 3: F0 mean given weight on the "low pitch" condition

3.5.2 Temporal Composition

To evaluate the effectiveness of temporal guidance, we combine the previous set of audio events pairs over time. For *ComposableART*, we use a *sigmoid*-like curve to increase and decrease over time the weights on event 1 and event 2 respectively. For the equivalent linguistic baseline, we create the instruction "Event 2 followed by Event 1". Figure 2 showcases time-windowed CLAP scores, y-axis, for each approach and event across time, x-axis. We observe that the baseline is unable to establish the temporal trend as well as *ComposableART*, where we observe, as expected, a consistent increase in event 1 and consistent decrease in event 2.

3.5.3 Task and Model Composition

Task Composition: In our website, we provide qualitative samples where we compose a set of tasks involving "electronic music", "birds chirping", "dog barking", and "TTS". Our results show that the output conforms to the composition of such tasks using the proposed method.

Model Composition: In our website, we provide samples where we consider 2 different *Fugatto* models, one trained on speech datasets and the other trained on general sounds and audio events. We perform model composition to synthesize samples that contain both speech as well as audio events. This can be immensely useful in the future, where each domain specific model is an expert model and a combination of such high-quality experts can be used to synthesize compositional outputs without the need to train a monolithic large generative model.

4 Discussion and Limitations

We work towards a future where unsupervised multitask learning in audio synthesis and transformation emerges from data and model scale. Our proposed framework *ComposableART Fugatto* establishes our first step towards this direction, and in this step we become aware of our current limitations and challenges. For example, optimizing dataset sampling weights to drive performance on multiple benchmarks is a herculean task, and generative models for audio and video would certainly benefit from research similar to (Albalak et al., 2023; Chung et al., 2023; Xie et al., 2024). Along straighter paths, we plan to replace mels with a latent representation that better supports low frequencies and stereo. We believe this modification should be rather straight forward. Further work is necessary to

establish the impact of data and free-form instructions on *Fugatto*'s emergent abilities, especially using language to combine tasks not jointly seen during training. Finally, *ComposableART* requires more analysis on the choice of weights, and how the norm of the vector field can be used for easier and more stable control.

Reproducibility Statement

We plan to release our dataset and code to facilitate reproducible research.

References

Alon Albalak, Liangming Pan, Colin Raffel, and William Yang Wang. Efficient online data mixing for language model pre-training. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*, 2023.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pp. 41–48, 2009.

Mathieu Bernard and Hadrien Titeux. Phonemizer: Text to phones transcription for multiple languages in python. *Journal of Open Source Software*, 6(68):3958, 2021.

Paul Boersma and Vincent Van Heuven. Speak and unspeak with praat. *Glot International*, 5(9/10): 341–347, 2001.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. *arXiv preprint arXiv:2304.09151*, 2023.

Shuqi Dai, Ming-Yu Liu, Rafael Valle, and Siddharth Gururani. Expressivesinger: Multilingual and multi-style score-based singing voice synthesis with expressive performance control. In *ACM Multimedia 2024*, 2024.

SeungHeon Doh, Keunwoo Choi, Jongpil Lee, and Juhan Nam. Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*, 2023.

Yilun Du, Shuang Li, and Igor Mordatch. Compositional visual generation with energy based models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6637–6647. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/49856ed476ad01fcff881d57e161d73f-Paper.pdf.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *International Conference on Machine Learning*, pp. 1068–1077. PMLR, 2017.

Sefik Emre Eskimez, Xiaofei Wang, Manthan Thakker, Canrun Li, Chung-Hsien Tsai, Zhen Xiao, Hemin Yang, Zirun Zhu, Min Tang, Xu Tan, et al. E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts. *arXiv preprint arXiv:2406.18009*, 2024.

Arushi Goel, Zhifeng Kong, Rafael Valle, and Bryan Catanzaro. Audio dialogues: Dialogues dataset for audio and music understanding. *arXiv preprint arXiv:2404.07616*, 2024.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*, 2023.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself, 2024. URL https://arxiv.org/abs/2406.02507.

Rafal Kocielnik, Shrimai Prabhumoye, Vivian Zhang, R Michael Alvarez, and Anima Anandkumar. Autobiastest: Controllable sentence generation for automated and open-ended social bias testing in language models. *arXiv preprint arXiv:2302.07371*, 2023.

Zhifeng Kong, Wei Ping, Ambrish Dantrey, and Bryan Catanzaro. Cleanunet 2: A hybrid speech denoising model on waveform and spectrogram. *arXiv preprint arXiv:2309.05975*, 2023.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. *arXiv preprint arXiv:2402.01831*, 2024a.

Zhifeng Kong, Sang-gil Lee, Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, Rafael Valle, Soujanya Poria, and Bryan Catanzaro. Improving text-to-audio models with synthetic captions. *arXiv preprint arXiv:2406.15487*, 2024b.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, et al. Voicebox: Text-guided multilingual universal speech generation at scale. *Advances in neural information processing systems*, 36, 2024.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=iTtGCMDEzS_.

Yeonghyeon Lee, Inmo Yeon, Juhan Nam, and Joon Son Chung. Voiceldm: Text-to-speech with environmental context. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12566–12571. IEEE, 2024.

Yichong Leng, Zhifang Guo, Kai Shen, Xu Tan, Zeqian Ju, Yanqing Liu, Yufei Liu, Dongchao Yang, Leying Zhang, Kaitao Song, et al. Prompttts 2: Describing and generating voices with text prompt. *arXiv preprint arXiv:2309.02285*, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Haohe Liu, Ke Chen, Qiao Tian, Wenwu Wang, and Mark D Plumbley. Audiosr: Versatile audio super-resolution at scale. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1076–1080. IEEE, 2024.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models, 2023. URL https://arxiv.org/abs/2206.01714.

Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Justin Lovelace, Soham Ray, Kwangyoun Kim, Kilian Q Weinberger, and Felix Wu. Simple-tts: End-to-end text-to-speech synthesis with latent diffusion. 2023.

Weili Nie, Arash Vahdat, and Anima Anandkumar. Controllable and compositional generation with latent-space energy-based models, 2021. URL https://arxiv.org/abs/2110.10873.

OpenAI. Gpt-4o: A powerful multimodal language model. https://openai.com/research/hello-gpt-4o, 2024. Accessed: 2024-09-21.

Ashis Pati, Siddharth Kumar Gururani, and Alexander Lerch. dmelodies: A music dataset for disentanglement learning. In *Proceedings of the 21th International Society for Music Information Retrieval Conference, ISMIR 2020, Montreal, Canada, October 11-16, 2020*, pp. 125–133, 2020. URL http://archives.ismir.net/ismir2020/paper/000300.pdf.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. The musdb18 corpus for music separation. 2017.

Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4779–4783. IEEE, 2018.

Spotify. Pedalboard documentation. `https://spotify.github.io/pedalboard/index.html`, 2024. Accessed: 2024-09-23.

Alexander Tong, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Kilian Fatras, Guy Wolf, and Yoshua Bengio. Conditional flow matching: Simulation-free dynamic optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Rafael Valle, Kevin Shih, Ryan Prenger, and Bryan Catanzaro. Flowtron: an autoregressive flow-based generative network for text-to-speech synthesis. *arXiv preprint arXiv:2005.05957*, 2020.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, et al. Audiobox: Unified audio generation with natural language prompts. *arXiv preprint arXiv:2312.15821*, 2023.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

Xiaofei Wang, Manthan Thakker, Zhuo Chen, Naoyuki Kanda, Sefik Emre Eskimez, Sanyuan Chen, Min Tang, Shujie Liu, Jinyu Li, and Takuya Yoshioka. Speechx: Neural codec language model as a versatile speech transformer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36, 2024.

Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. Understanding and improving layer normalization. *Advances in neural information processing systems*, 32, 2019.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022.

Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, et al. Uniaudio: An audio foundation model toward universal audio generation. *arXiv preprint arXiv:2310.00704*, 2023.

Jinhyeok Yang, Junhyeok Lee, Hyeong-Seok Choi, Seunghun Ji, Hyeongju Kim, and Juheon Lee. Dualspeech: Enhancing speaker-fidelity and text-intelligibility through dual classifier-free guidance. *arXiv preprint arXiv:2408.14423*, 2024.

A Appendix

A.1 Datasets

In this section we provide information about our dataset generation strategy, tasks, instruction generators, and a full list of datasets created and used during training, including their sampling weights and task probabilities.

A.1.1 Approach to Dataset Generation

In this subsection we provide further descriptions of the procedures in each of the five pillars described in Section 2.1. Overall, the captions and instructions generation process is centered around leveraging LLMs to transform tags into richer descriptions, always being careful that the description relates to the audio content, and to create what we call *instruction generators*, i.e. python methods that create free-form instructions based on the synthetic descriptions and task at hand.

I – The python code snippet below is an LLM generated script that outputs free-form instructions for the task *reverse sound*, for example. Once such a script exists, it can be used to prompt an LLM to generate scripts for other tasks with increasing levels of complexity.

```
class ReverseAudioInstructor:
    audio_references = {
        'standard': {
            'verbs': ['reverse', 'play backward', 'invert'],
            'gerunds': ['reversing', 'playing backward', 'inverting'],
            'contexts': ['audio', 'sound', 'recording', 'clip', 'track'],
            'asks': ['Can you', 'Please', 'Could you', 'I need you to'],
            'styles': ['completely', 'precisely', 'accurately'],
            'mentions': ['I provided', 'I sent', 'I attached']
        },
        ...
    }

    @staticmethod
    def generate_instruction(persona='standard'):
        ref = ReverseAudioInstructor.audio_references[persona]

        templates = [
            "{ask} {verb} the {context}.",
            "{verb} the {context} {style}.",
            "{ask} {verb} the {context} {mention}.",
            "Your task is to {verb} the {context}.",
            "We need the {context} {mention} to be {gerund}.",
            "{gerund} the {context} is the goal.",
            "Please focus on {gerund} the {context}.",
            "The objective is {gerund} the {context} {mention}.",
        ]

        template = random.choice(templates)

        instruction = template.format(
            ask=random.choice(ref['asks']),
            verb=random.choice(ref['verbs']),
            gerund=random.choice(ref['gerunds']),
            context=random.choice(ref['contexts']),
            style=random.choice(ref['styles']),
            mention=random.choice(ref['mentions'])
        )

        return instruction.capitalize()


if __name__ == '__main__':
    for persona in ReverseAudioInstructor.audio_references.keys():
        print(f"\n{persona.capitalize()} instructions:")
```

II – The python code snippet below is an LLM generated script that outputs free-form instructions for the task *speech modulation*, for example.. Note that the instructions refer to relative changes such as increase and decrease with different magnitudes such small or large changes. Once such a script exists, it can be used to prompt an LLM to generate scripts for other tasks with increasing levels of complexity.

```
class SpeechModulationInstructor:
    instructions = {
        'scale_formant': {
            'increase': {
                'small': [
                    "add a touch more resonance",
                    "slightly enhance the formant frequencies",
                    ...
                ],
                ...
                'large': [
                    "dramatically enhance the resonance",
                    "massively boost the formant frequencies",
                    ...
                ]
            },
            'decrease': {
                'small': [
                    "tone down the resonance just a little",
                    "slightly reduce the formant frequencies",
                    ...
                ],
                ...
                'large': [
                    "dramatically reduce the resonance",
                    "massively lower the formant frequencies",
                    ...
                ]
            }
        },
        ...
        }
    }

    @staticmethod
    def get_instruction(modulation, direction, intensity):
        return random.choices(SpeechModulationInstructor.instructions[modulation][direction][intensity])[0]

    @staticmethod
    def combine_instructions(modulations):
        parts = []
        for modulation_i in modulations:
            modulation = modulation_i['modulation']
            direction = modulation_i['direction']
            intensity = modulation_i['intensity']
            instruction_part = SpeechModulationInstructor.get_instruction(modulation, direction, intensity)
            parts.append(instruction_part)
        instruction = ""
        if parts:
            # Combine parts into a single instruction
            if len(parts) > 1:
                combined_instruction = ", and ".join(parts[:-1]) + ", and " + parts[-1]
            else:
                combined_instruction = parts[0]
            instruction = "Let's " + combined_instruction + "."
        return instruction

if __name__ == '__main__':
    print("main")
```

III - Figure 4 provides a visual depiction of our speech captioning, dubbed Prompt-2-Voice (P2V) pipeline. We applied this approach to several open-source soeech datasets. Additionally, we incorporated existing speaker descriptions from datasets like PromptSpeech (Guo et al., 2023), which provide details on gender, pitch, volume, and speaking rate, and enriched them by adding emotion and quality information. This automation not only streamlines the process but also allows us to efficiently label in-the-wild datasets, significantly scaling the available training data wigh high quality captions.

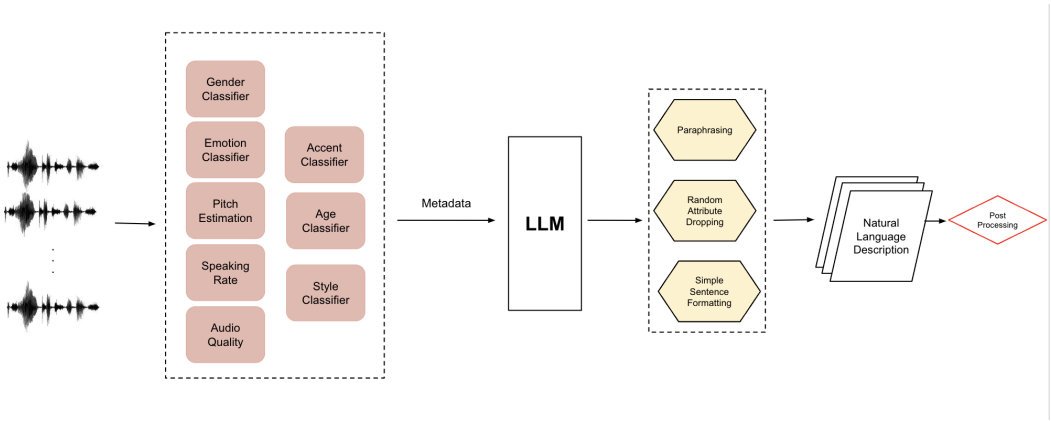


Figure 4: Synthetic caption generation pipeline for Prompt-to-Voice (P2V).

V - We use audio effects libraries to create synthetic paired data where some factors are held constant while others change. We use Pedalboard (Spotify, 2024), a Python library for audio effects, to apply different audio effects to sounds believed not to have effects. For each effect, there are a number of parameters that can be altered. We generate on the fly audio segments with the same effect but different levels of a single parameter while keeping the others fixed, so that we can not only create instructions with synthetic pairs of uneffected vs. effected data, but also synthetic pairs where we ask the model to increase or decrease the intensity of a given parameter (e.g. ”increase the compression rate a little bit” or ”reduce the room size for the reverb moderately”). For each parameter, we determined a reasonable range for the parameter, and increments that would correspond to ”a little”, ”moderate”, and ”a lot” of that parameter.

A.1.2 List of datasets

Open Source Vocal Datasets

Table 7: Open source vocal datasets. tts is Text-To-Speech and tta is Text-To-Audio (Audio from Caption)

Dataset	Sampling Weights	Task and Probabilities
AISHELL-3	2.46	tts
CML-Dutch	0.769	tts
CML-French	0.334	tts
CML-German	1.762	tts
CML-Italian	0.158	tts
CML-Polish	0.051	tts
CML-Portuguese	0.086	tts
CML-Spanish	0.512	tts
common-accent-AccentClassification	0.131	tta
CREMA-D-EmotionClassification	0.074	tta
DAPS-Enhancement	0.21	enhancement paired 0.99 deenhancement paired 0.01
DNS-Challenge-2020	14.247	source separation
emov-db-EmotionClassification	0.068	tta
IEMOCAP-EmotionClassification	0.059	tta
jl-corpus-EmotionClassification	0.024	tta
LibriTTS-Clean-100	1.23	tts
LibriTTS-Clean-360	4.31	tts
LibriTTS-Other-500	0.643	tts
LibriVox-English	52.54	tts
LibriVox-French	0.909	tts
LibriVox-German	1.147	tts
LibriVox-Italian	0.114	tts
LibriVox-Portuguese	0.12	tts
LibriVox-Spanish	0.276	tts
LIMMITS2024-*	0.768	tts 0.80 inpainting 0.10 upsampling 0.10
MSP-PODCAST-Publish-1.9-EmotionClassification	0.451	tta
NonSpeech7k-EventClassification	0.063	tta
OMGEmotion-EmotionClassification	0.017	tta
ravdess-EmotionClassification	0.014	tta
SongDescriber-AudioCaptioning	0.077	tta
SONYC-UST-EventClassification	0.279	tta
tess-EmotionClassification	0.028	tta
VCTK-VoiceConversion	2.89	voice conversion paired vctk
VCTK-TTS	0.137	tts
VocalSound-VocalClassification	0.155	tta

Open Source Non-Vocal Datasets

Table 8: Open Source Non-Vocal Datasets. tts is Text-To-Speech and tta is Text-To-Audio (tta)

Dataset	Sampling Weights	Task and Probabilities
audiocaps-AudioCaptioning	5.21	tta 0.90 inpainting 0.10
BBCSoundEffects-AudioDescription	0.15	tta
chime-home-EventClassification	0.05	tta
Clotho-AQA-EventClassification	0.01	tta
Clotho-AQA singlelabel-EventClassification	0.07	tta
Clotho-v2-AudioCaptioning	0.19	tta
CochlScene-SceneClassification	0.61	tta
Epidemic sound-AudioCaptioning	0.41	tta
ESC-50	1.12	tta 0.90 inpainting 0.10
FMA-GenreClassification	1.04	tta
FSD50k-EventClassification	0.41	tta
GTZAN-GenreClassification	0.01	tta
LP-MusicCaps-MC-AudioCaptioning	0.07	tta
LP-MusicCaps-MSD-0	8.56	tta 0.90 inpainting 0.10
LP-MusicCaps-MSD-1	8.62	tta 0.90 inpainting 0.10
LP-MusicCaps-MSD-AudioCaptioning	11.82	tta
LP-MusicCaps-MTT-AudioCaptioning	0.47	tta
MACS-AudioCaptioning	0.17	tta
Maestro	0.08	tta 0.90 upsampling 0.10
Medley-solos-DB	0.19	tta 0.90 upsampling 0.10
MSD	12.38	tta 0.80 upsampling 0.10 inpainting 0.10
MTG-Jamendo	2.16	tta 0.90 inpainting 0.10
musdbhq-InstrClassification	0.10	tta
MusicCaps-AudioCaptioning	0.54	tta 0.90 inpainting 0.10
NonSpeech7k-EventClassification	0.06	tta
NSynth-MIR	2.31	tta
SongDescriber-AudioCaptioning	0.08	tta
SONYC-UST-EventClassification	0.28	tta
SoundDescs-AudioDescription	0.23	tta
SoundVE-Caps	37.00	tta
UrbanSound8K-EventClassification	0.09	tta
WavText5K-AudioCaptioning	0.04	tta
WavText5K-Tagging	0.02	tta

New Datasets generated from Open Source Data

Table 9: New datasets generated from open source data. tts is Text-To-Speech, tta is Text-To-Audio (Audio from Caption), and P2V refers to prompt to voice, in which LLMs were used to create full form captions given speech attributes. The -AF suffix indicates synthetic captions generated with Audio Flamingo.

Dataset	Sampling Weights	Task and Probabilities
audiocaps-AudioCaptioning-AF	5.79	tta 0.90 inpainting 0.10
AudioSet-AF	37.24	tta 0.80 inpainting 0.05 inpainting random mask 0.05 upsampling 0.09 downsampling 0.01
AISHELL-3-AddRemove-Sound-Effects	3.23	add sound effects 0.50 remove sound effects 0.50
AISHELL-3-SoundEffectsModulation	3.84	sound effects modulation
AISHELL-3-SpeechModulationPraat	3.84	speech modulation praat
CLAP freesound-AF	2.82	tta
CREMA-D-P2V	0.62	tts 0.80 inpainting 0.10 upsampling 0.10
EGFxSet-AddSoundEffects	0.00	add sound effects paired
EGFxSet-RemoveSoundEffects	0.00	remove sound effects paired
Emilia [English] (subset + additions)	27.23	tts 0.80 inpainting 0.05 inpainting random mask 0.05 upsampling 0.04 reverse sound 0.04 downsampling 0.01
emov-db-EmotionClassification	0.07	tta
ESD-ChangeEmotion	1.90	speech modulation paired
ESD-ENGLISH-P2V	0.93	tts 0.80 inpainting 0.10 upsampling 0.10
ESD-MANDARIN-P2V	0.63	tts 0.80 inpainting 0.10 upsampling 0.10
IEMOCAP-EmotionClassification	0.06	tta
jl-corpus-EmotionClassification	0.02	tta
LibriTTS-Clean-100-Add-Remove-Sound-Effects	7.85	add sound effects 0.50 remove sound effects 0.50
LibriTTS-Clean-100-SoundEffectsModulation	0.06	sound effects modulation
LibriTTS-Clean-100-SpeechModulationPraat	9.34	speech modulation praat
LibriTTS-Clean-100-Enhancement	0.85	enhancement paired 0.99 deenhancement paired 0.01
LibriTTS-Clean-360-Enhancement	2.71	enhancement paired 0.99 deenhancement paired 0.01
LibriTTS-Other-500-Enhancement	6.11	enhancement paired
JL-CORPUS-P2V	0.62	tts 0.80 inpainting 0.10 upsampling 0.10
LMD-Aligned	1.17	midi2audio
musdbhq-InstrClassification	0.10	tta
musdbhq-add-sound	3.81	add sound 0.50 add sound to mixture 0.50
musdbhq-singing	0.86	singing voice synthesis
musdbhq-singing-aggregated	0.86	singing voice synthesis
musdbhq-source-separation	3.81	source separation 0.50 remove sound from mixture 0.50
NonSpeech7k-EventClassification	0.06	tta
NSynth-MIR	2.31	tta
OMGEmotion-EmotionClassification	0.02	tta
ravdess-EmotionClassification	0.01	tta
RAVDESS-CreateVariation	0.02	speech modulation paired
RAVDESS-ChangeIntensity	0.15	speech modulation paired
RAVDESS-ChangeEmotion	0.22	speech modulation paired
RAVDESS-P2V	0.62	tts 0.80 inpainting 0.10 upsampling 0.10
ExpressiveSinger-P2V	2.84	singing voice synthesis nolanguage
tess-EmotionClassification	0.03	tta
VGG-AF	0.92	tta
WavCaps-AF	7.83	tta

Below we provide descriptions for the suffixes associated with our generated datasets presented in Table 7, Table 8, and Table 9.

–AF: Refers to generating synthetic captions with the strategy described in (Kong et al., 2024b). In summary, the strategy consists of using an audio understanding model, here Audio Flamingo Chat, to caption sounds in the wild, and then filtering out synthetic captions based on the CLAP cosine similarity between the audio and the synthetic caption. In this paper, we used this strategy to create synthetic captions for AudioCaps, AudioSet, CLAP Freesound, and WavCaps.

–{Add, Remove} Sound Effects: These refer to applying, on the fly, audio effects modifications to existing audio data to create paired data that describe tasks related to adding and removing sound effects. We focus on speech, which we assume has the least amount of audio effects applied to it, especially when compared to music data. In this iteration, the audio modulations are performed with (Spotify, 2024) and include a large list of effects such as chorus, reverb, distortion, amongst others.

–{Add, Remove}: This refers to splitting an audio mixture into separate audio stems and creating manifests that add one track given another track. In this *opus*, we have not explored creating artificial mixtures by adding random waveforms but imagine this strategy can yield good results.

–P2V: P2V, or prompt-to-voice, is a task that enables control of speech synthesis through textual prompts, allowing for the description of speaker characteristics when an appropriate audio prompt that matches the desired persona is unavailable. The strategy has several components. First, we extract and curate all possible tags from existing metadata. For example, the expressive singer (Dai et al., 2024) dataset includes, implicitly and explicitly, information about vocal range, accent, language, style and others. Then, following the strategy in Section 2.1 we first prompt LLMs to produce long form descriptions of each attribute, then we prompt an LLM to create a instruction generator that combines the modified attributes in different ways.

–Singing: refers to leveraging music stems dataset with vocal tracks to create singing datasets. As usual, we leverage all the metadata available, combined with transcriptions obtained from speech transcription models and song captions obtained by prompting LLMs. In cases where an accompaniment or backing track is available, we associate the timestamps on each lyrics snippet with the respective accompaniment. Once the metadata types are available, we prompt an LLM to create a python method that produces instructions given the metadata for a singing voice synthesis task with or without accompaniment.

–Singing-Aggregated: -Singing-Aggregated adds to -Singing the combination of adjacent sentences, given criteria such as max gap between sentences and max sentence length.

–Source Separation and Add To Mixture: refers to leveraging existing In this iteration, assuming doing such would produce samples undesirable out-of-distribution samples, we do not create mixtures by combining random samples together. Instead, we leverage mixtures that are already exist.

–Speech Modulation: refers to applying speech modulations to existing speech data to create paired data that describe a task in which a relative change is being applied to an attribute of speech, e.g. "*Increase* the speaking rate in this sample". In this iteration, the speech modulations are performed with Praat (Boersma & Van Heuven, 2001) and we focus on transformations such as formant scaling, F0 mean scaling (equivalent to transposition in music), F0 variance scaling (equivalent to flattening or expanding the F0 contour), and speaking rate scaling (equivalent to making a person speak faster or slower). Due to artifacts that can be introduced in the audio as a result of the modulation, we limit the scaling values to ranges that we find acceptable in terms of audio quality.

–Speech Modulation Paired: This refers to leveraging available paired data to create new tasks. For example, an emotional speech dataset with paired data can be used to enable an emotion transformation tasks. The procedure is straight forward and consists of first grouping samples by speaker and transcript, then creating pairs that establish relationships between two samples. For example, given a speaker and transcript, two samples with different emotion can be used to define a "convert emotion task", two samples with the same emotion and intensity can be used to create a "create a variation of this speech", and two samples with the same emotion but different intensity can be used to define a "increase the intensity in the emotion task". Once the pairs and new tasks are defined, we prompt an LLM to create a script that takes in the attributes and tasks to generate task-specific instructions.

–**Sound Effects Modulation:** refers to applying sound effects modulations to existing data to create paired data for relative change in audio effects in a file. In this iteration, the speech modulations are performed with Praat (Boersma & Van Heuven, 2001) and focus on formant scaling, F0 mean scaling, F0 variance scaling and speaking rate scaling. We limit the values to the range below

A.1.3 TASKS

Below we provide the list of tasks that are known to be supported by our model, and the respective minimal instruction for that task. Please note that we provide minimal instructions due to layout reasons, not due to model limitations.

Task	Minimal Instruction
Add Sounds to Instrument Track	
using audio example	add drums like this `<audio>` to this guitar track `<audio>`
using text description	add rock drums to this guitar track `<audio>`
Add Sounds to Sound Mixture	
using audio example	add guitar like this `<audio>` to this backing track `<audio>`
using text description	add guitar to this backing track `<audio>`
Apply Sound Effects	add long reverb to this audio `<audio>`
Copy Audio	copy this `<audio>`
Downsample Audio	downsample this to 16kHz `<audio>`
Enhance Audio	enhance this sound `<audio>`
Generate Audio From Captions	generate a saxophone barking
Generate Music From Captions	generate a calm sound track
Generate Speech From Captions	generate an angry voice
Inpaint Audio (Continuation)	inpaint this sound `<audio>`
MIDI2Audio	
using audio example	turn this MIDI track `<audio>` into natural audio.
using text description	turn this MIDI track `<audio>` into a heavy metal track.
Remove Sound Effects	remove sound effects `<audio>`
Remove Sound from Sound Mixture	
from audio examples	remove this `<audio>` from these sounds
from text captions	remove the piano from this music track `<audio>`
Reverse Sound	reverse this sound
Singing-Voice Synthesis	
given Speech Captions	sing this 'At last my love' with a male voice
given Language	sing this 'At last my love' in English
given Accent	sing this 'At last my love' in English with an Italian accent
Separate Audio into Sources	
given audio examples	give me this `<audio>` from this music track `<audio>`
given text description	give me the piano from this music track `<audio>`
Sound Effects Modulation	(Increase and Decrease Attributes)
Chorus	increase the chorus a bit `<audio>`
Compressor	decrease the compressor threshold `<audio>`
Delay	decrease the delay time `<audio>`
Distortion	increase the distortion `<audio>`
Limiter	make the limiter less strong `<audio>`
Phaser	increase the phaser speed `<audio>`
Reverb	increase the reverb room size in this `<audio>`
Speech Modulation (Paired)	
Voice Conversion	convert from this `<audio>` to this `<audio>`
Accent Conversion	convert from British English to American English
Emotion Conversion	make this calm sample `<audio>` sound angry
Speech Variation	give me a different take on this voice `<audio>`
Speech Modulation (Praat)	
Scale Speaking Rate	increase the speaking rate `<audio>`
Scale F0 Mean	increase the average pitch `<audio>`
Scale F0 Variance	increase the variance in pitch `<audio>`
Scale Formants	scale the formants here `<audio>`
Text-To-Speech	
from Speech Prompt	say this 'May the force!' given this voice `<audio>`
from Speech Captions	say this 'May the force!' with a male voice
from Language	say this 'May the force!' in English
from Accent	say this 'May the force!' in English with an Italian accent
Upsample Audio	upsample this sound `<audio>`

A.1.4 INSTRUCTIONS

After an informal evaluation, we concluded that, for our purpose, Claude Sonnet is better than GPT4-o at producing instructions and following prompts. Below we provide examples of template-based and free-form instructions for a handful of tasks and datasets. For clarity, we remove the redundant 'input:' and 'output:' parts from all instructions.

AISHELL-3-AddRemove-Sound-Effects
```
Eradicating the audio from the provided audio material with the
Distortion, and Phaser effect is your task.  Focus on eradicating
it systematically.
```

AISHELL-3-SoundEffectsModulation
```
We need to minimize the delay time.</caption>
```

AISHELL-3-SpeechModulationPraat
```
Let's subtly enhance the pitch for a brighter sound, and
dramatically slow down the speech for a very relaxed
delivery.</caption>
```

audiocaps-AudioCaptioning
```
synthesize A consistent, loud mechanical motor</caption>
```

AudioSetFullwoAudioMusicCaps-EventClassification
```
synthesize This is a sound of Speech</caption>
```

AudioSet-AF
```
Yo, mind fill in the missing bits in this tune?  Please do this
smoothly.]
```

CREMA-D-P2V
```
Manifest a voice reproduction in American English verbalizing "The
airplane is almost full.", with the timbre is often monotone and
lacks energy, and itś like a middle-aged who is not Hispanic.
```

LibriTTS-Clean-100
```
Stitch up a spiel in en declaring ""But there was a passenger
dropped off for you-a little girl.", with a female speaker with
a moderate pitch and intonation delivers her words quite rapidly
in a confined, clear acoustic environment.
```

RAVDESS-ChangeEmotion
```
modulate I want to switch from angry to fearful.</caption> given
example:
```

A.2 Model and Training

Model visualization: Figure 5 provides a visual description of *Fugatto*'s architecture and input handling and Algorithm 1 provides a pseudo-algorithm for the optimal-transport conditional flow matching loss. Essentially, it minimizes the mean squared error between the estimator's prediction and a linear interpolation between the data and the Gaussian noise sample used to condition the model on x_t, scaled by $(1 - \sigma)$, where σ is a small enough value.

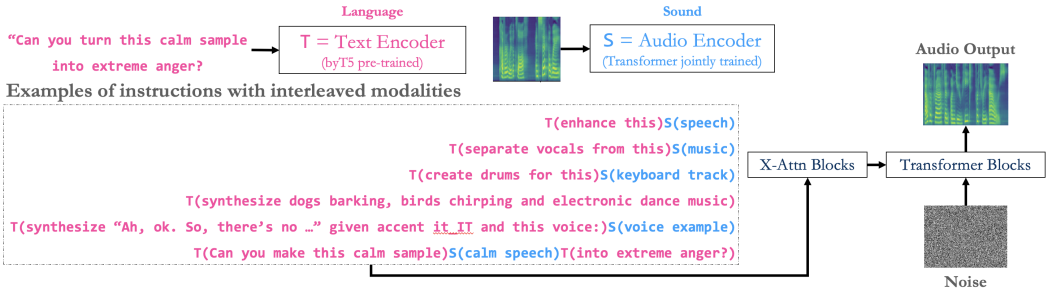


Figure 5: A description of *Fugatto*'s architecture and input handling.

Algorithm 1 Optimal Transport Conditional Flow Matching Loss Pseudo-Algorithm

1: Sample $\mathbf{x}_1 \sim \mathcal{X}$, where $\mathcal{X}$ is the data distribution
2: $\sigma \leftarrow 0.01$
3: $\mathbf{x}_0 \leftarrow \texttt{randn\_like}(\mathbf{x}_1)$
4: $\mathbf{x}_t \leftarrow (1 - (1 - \sigma) \cdot t) \cdot \mathbf{x}_0 + t \cdot \mathbf{x}_1$
5: $\mathbf{v}_t \leftarrow estimator_\theta(\mathbf{x}_t, t;)$
6: $\mathbf{u}_t \leftarrow \mathbf{x}_1 - \mathbf{x}_0 \cdot (1 - \sigma)$
7: loss $\leftarrow \texttt{mse}(\mathbf{v}_t, \mathbf{u}_t)$

Optimal Transport Conditional Flow Matching Pseudo-Algorithm:

Model hyperparameters: We provide a list of *Fugatto*'s hyperparameters in Table 10.

Table 10: Model Hyperparameters for the main *Fugatto* evaluated in this paper.

Hyperparameter	Value
t_schedule	uniform
n_mel_channels	80
n_hidden	1536
sigma	0.01
text_encoder_config.name	google/byt5-large
text_encoder_config.scale	1.0
text_encoder_config.n_hidden	1536
mel_encoding_strategy	separate
mel_encoder_config.is_causal	false
mel_encoder_config.pos_emb.name	rope
mel_encoder_config.pos_emb.base	16384
mel_encoder_config.use_flash_attention	true
mel_encoder_config.deterministic	false
mel_encoder_config.n_layers	3
mel_encoder_config.p_dropout	0.1
mel_encoder_config.p_dropout_out	0.0
mel_encoder_config.n_heads	16
mel_encoder_config.has_xattn	false
mel_encoder_config.apply_norm_to_cond	false
mel_encoder_config.layer_norm_method	pre
mel_encoder_config.kernel_size	3
mel_encoder_config.use_layer_scale	true
mel_encoder_config.layer_scale_init	0.1
mel_encoder_config.layer_scale_decay	0.95
mel_encoder_config.d_model	1536
decoder_config.d_time	128
decoder_config.transformer_hparams.is_causal	false
decoder_config.transformer_hparams.pos_emb.name	rope
decoder_config.transformer_hparams.pos_emb.base	16384
decoder_config.transformer_hparams.use_flash_attention	true
decoder_config.transformer_hparams.deterministic	false
decoder_config.transformer_hparams.n_layers	24
decoder_config.transformer_hparams.p_dropout	0.1
decoder_config.transformer_hparams.n_heads	16
decoder_config.transformer_hparams.has_xattn	true
decoder_config.transformer_hparams.kernel_size	3
decoder_config.transformer_hparams.context_xattn.n_heads	16
decoder_config.transformer_hparams.context_xattn.d_heads	1536
decoder_config.d_data	80
decoder_config.d_model	1536

Training and Inference During the first phase, *Fugatto* is trained on at least 32 NVIDIA A100 GPU for approximately 1M iterations with template-based instructions and a subset of tasks. During the second phase, we restart the optimizer and train for approximately 250k iterations, sampling from template-based and free-form instructions uniformly and adding all tasks. We use the AdamW optimizer (Loshchilov & Hutter, 2019) with a learning rate of 1e-4, annealing the learning rate to 1e-6 during the second phase. A G2P model (Bernard & Titeux, 2021) pre-processes the text into the International Phonetic Alphabet (IPA) format. During inference, we generate mel-spectrograms using 50 function evaluations, 100 in practice, with Heun's Solver and task-specific guidance scale γ. Mel-spectrogram to waveform conversion is performed using the pre-trained universal BigVGAN V2 vocoder, available in the BigVGAN (Lee et al., 2023) repository[4].

[4]BigVGAN: https://github.com/nvidia/bigvgan

A.3 ABLATIONS

t-sampling We draw inspiration from (Shen et al., 2018; Valle et al., 2020), where, due to optimization issues, the model would be stuck in a local minima and not make use of the text conditioning variable, rendering the model useless during inference. We believe the same is true with $t \sim \text{sigmoid}(\mathcal{N}(0, 1))$, and that with such a distribution the model is stuck in a local minima and does not leverage the text to minimize the loss.

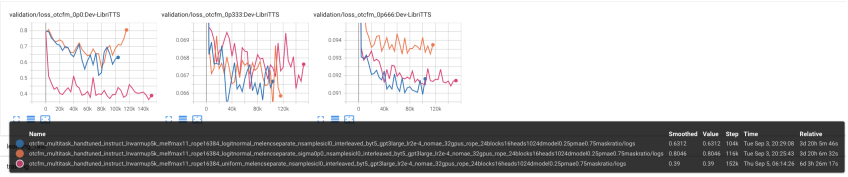


Figure 6: A description of the image.

Model capacity: Exponentially smoothed validation loss per task at different t-values from smaller models with 0.8 B, 1.4 B params, to larger models with 2.5 B parameters.

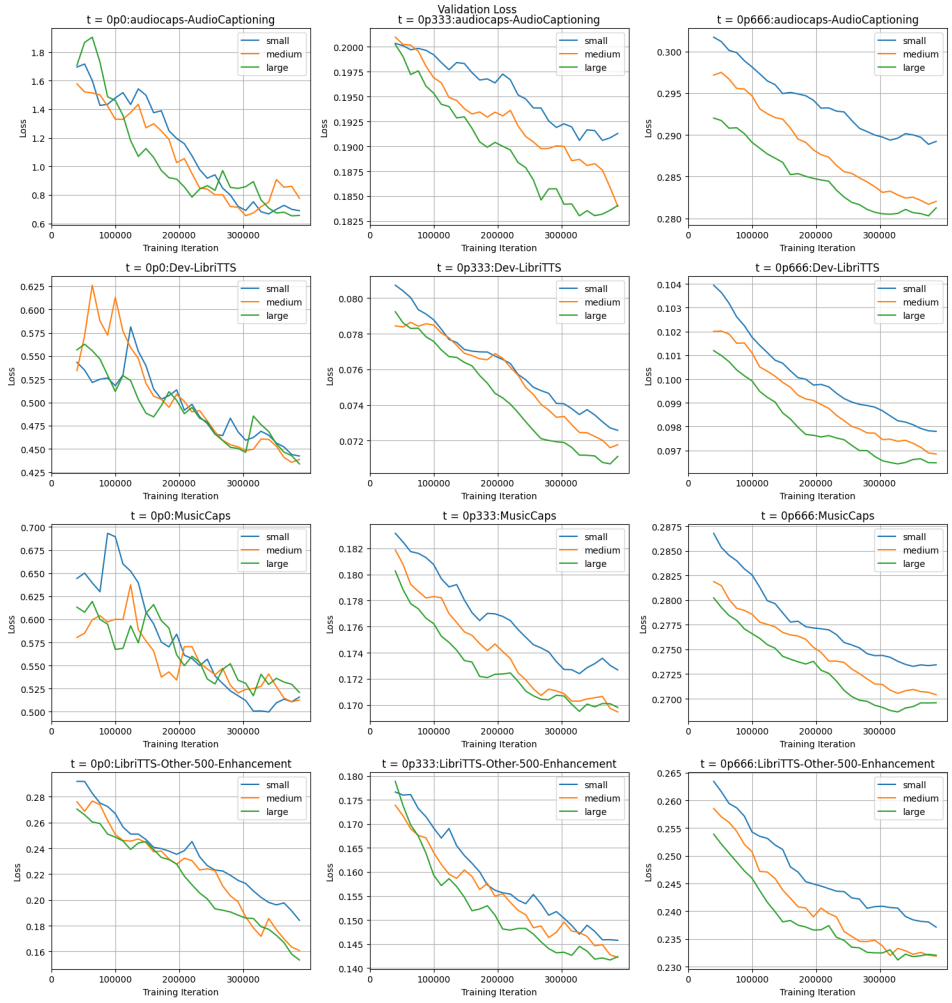


Figure 7: Validation scores for different *Fugatto* sizes on 3 benchmarks at 3 different t-values.

A.4 Singing Voice Synthesis Music Styles and Lyrics

The Singing-Voice-Synthesis experiments in Section 3.2 evaluates all combinations between the 13 lyrics snippets and 10 music styles below. For each combination, we use the singing-voice-synthesis instruction generator to create instructions such as: "Showcases a female singer with an interesting sound, conveys the message through american english lyrics, and infuses country influences throughout."

This is the set of 13 lyrics snippets used during evaluation:

```
"Is this the real life?\nIs this just fantasy?"
"As I walk through the valley of the shadow of death"
"Somebody once told me\nThe world is gonna roll me."
"Look\nIf you had\nOne shot\nOr one opportunity."
"Joy to the world\nThe lord is come."
"Carry on, my wayward son\nThere'll be peace when you are done."
"Please allow me to introduce myself."
"At first I was afraid, I was petrified."
"The world was on fire and no one could save me but you."
"Josie's on a vacation far away."
"She's got a smile that it seems to me."
"She was a fast machine, she kept her motor clean."
"Do you have the time to listen to me whine."
```

This is the set of 10 music styles used during evaluation:

```
"Country",
"Electronic",
"Hard Rock",
"Hip-Hop",
"Latin Rock",
"Metal",
"Opera",
"Pop",
"R&B",
"Singer-Songwriter"
```

A.5 MIDI2AUDIO F0 CONTOURS

Figure 8 shows comparisons of 25 monophonic melodies from the test set. As evident from the plot, the model manages to follow the provided MIDI notes and timing well, while inserting nuances, and varying timbres to the performance.

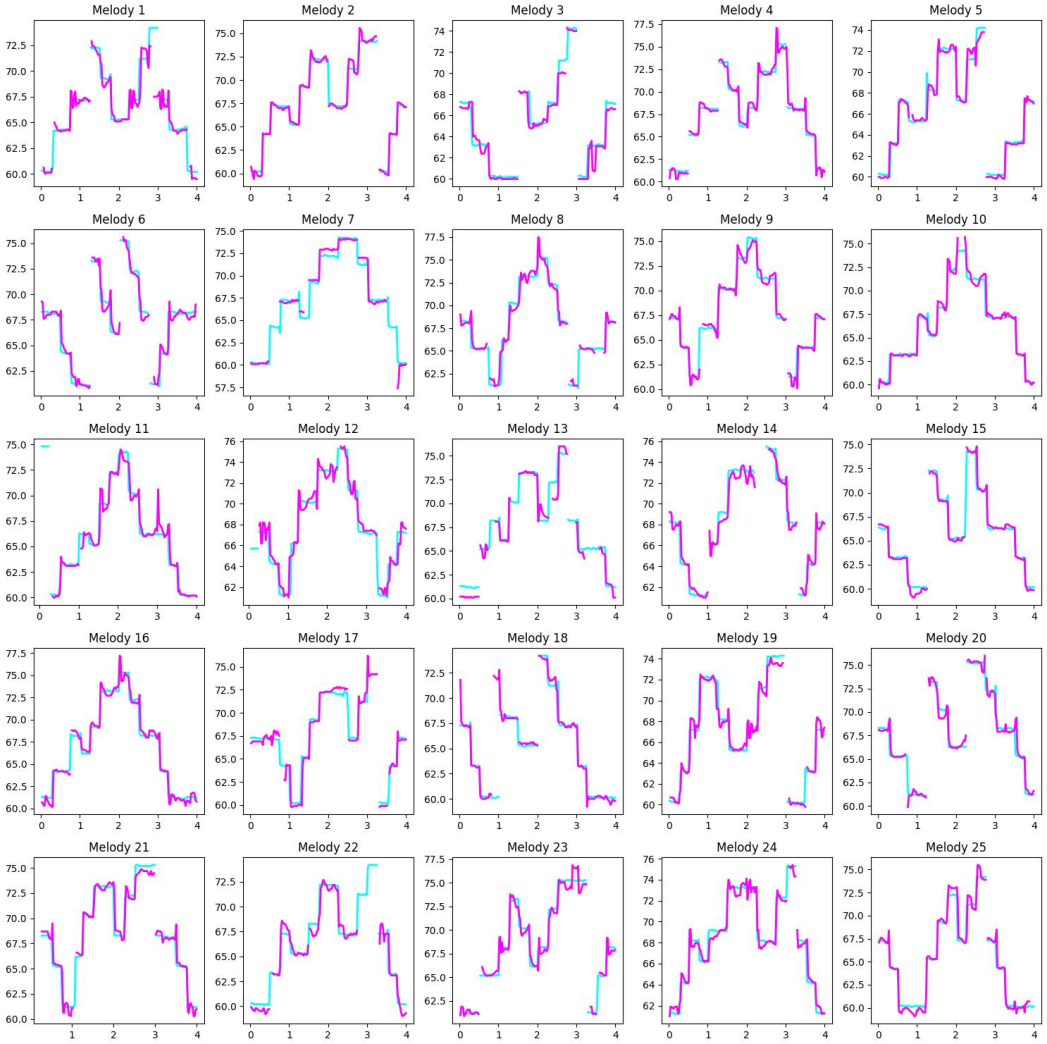


Figure 8: F0 contours of input MIDI (cyan), and generated melodies (magenta). X-axis denotes time in seconds, and Y-axis denotes MIDI pitch.

A.6 COMPOSITIONALITY

Attribute Composition (Audio Events): This is the list of audio events used during Attribute/Event composition in 3.5:

```
 Violin, fiddle; Accelerating, revving, vroom; Water; Acoustic
guitar; Afrobeat; Whistle; Air conditioning; Air horn, truck horn;
Aircraft; Wind
```

Exhibit C

Audio Flamingo: A Novel Audio Language Model with Few-Shot Learning and Dialogue Abilities

Zhifeng Kong [1] **Arushi Goel** [1] **Rohan Badlani** [1] **Wei Ping** [1] **Rafael Valle** [1] **Bryan Catanzaro** [1]

Abstract

Augmenting large language models (LLMs) to understand audio – including non-speech sounds and non-verbal speech – is critically important for diverse real-world applications of LLMs. In this paper, we propose **Audio Flamingo**, a novel audio language model with *1)* strong audio understanding abilities, *2)* the ability to quickly adapt to unseen tasks via in-context learning and retrieval, and *3)* strong multi-turn dialogue abilities. We introduce a series of training techniques, architecture design, and data strategies to enhance our model with these abilities. Extensive evaluations across various audio understanding tasks confirm the efficacy of our method, setting new state-of-the-art benchmarks. Our demo website is `https://audioflamingo.github.io/` and the code is open-sourced at `https://github.com/NVIDIA/audio-flamingo`.

1. Introduction

The ability to understand sound is unarguably important and necessary for an agent to interact with the world. While large language models (LLMs) have shown an impressive ability to understand and reason about the world through text, their understanding of sound remains limited to transcriptions of speech (Lyu et al., 2023), thus making LLMs agnostic to important information in non-speech sounds and non-verbal speech. Even though recent contributions have improved their ability to understand sound (Gong et al., 2023c; Lyu et al., 2023; Huang et al., 2023; Deshmukh et al., 2023; Chu et al., 2023; Tang et al., 2023a), there exists no model that combines: *i)* strong audio understanding ability on various tasks (Deshmukh et al., 2023), *ii)* the ability to execute multi-turn dialogues (Duan et al., 2023), and *iii)* the ability to quickly adapt to unseen tasks without fine-tuning, for example, through in-context learning (Alayrac et al., 2022) and retrieval augmented generation (Lewis et al., 2020).

Our contribution to expand LLM's ability to understand sound is called **Audio Flamingo**: a novel audio language model that supports in-context learning (ICL), retrieval augmented generation (RAG), and multi-turn dialogues. It achieves state-of-the-art results on multiple audio understanding tasks.

Expanding LLM's ability to understand sound is a challenging task. The first challenge is extracting features from variable-length audio, and conditioning the LM on the audio features. While prior works have designed representations for audio of any length (Wu et al., 2023), they can lose temporal information. In this work, we introduce an audio feature extractor with sliding window based on Elizalde et al. (2023b), which we believe to capture temporal information better. To condition the LM on audio inputs, previous models prepended language tokens with audio tokens (Deshmukh et al., 2023; Chu et al., 2023; Tang et al., 2023a). This approach may have excessive overhead especially for long audio, as the complexity is quadratic with respect to the number of audio tokens. In contrast, we use cross attentions to fuse audio inputs into the LM in a similar way as Flamingo (Alayrac et al., 2022). In our approach the complexity is linear in the number of audio tokens, thus making Audio Flamingo able to generalize to many audio inputs efficiently.

The second challenge is collecting and training on highly heterogeneous data. Prior works have collected and combined different datasets for training (Doh et al., 2023; Deshmukh et al., 2023; Chu et al., 2023; Gong et al., 2023c). We follow their approach and curate a heterogeneous dataset with approximately 5.9 million audio-text pairs. Prior works have also designed different training curriculum, such as training on close-ended tasks first followed by open-ended tasks (Gong et al., 2023c; Tang et al., 2023a). However, these result in a

[1]NVIDIA, CA, USA. Correspondence to: Zhifeng Kong `<zkong@nvidia.com>`, Wei Ping `<wping@nvidia.com>`, Rafael Valle `<rafaelvalle@nvidia.com>`.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

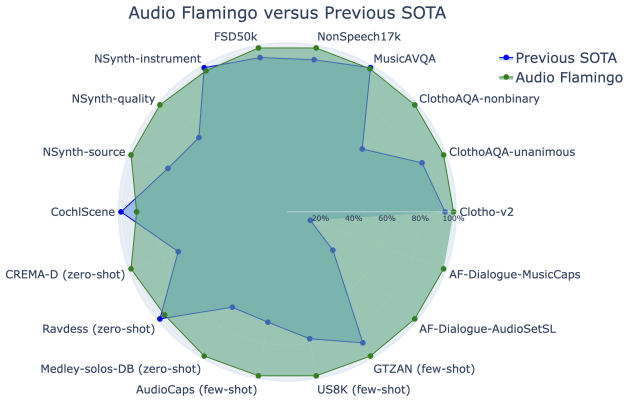


Figure 1. Audio Flamingo versus previous state-of-the-art (Deshmukh et al., 2023; Chu et al., 2023; Gong et al., 2023b;c; Tang et al., 2023a; Ghosh et al., 2023) on a number of audio understanding benchmarks. The numbers are normalized such that the maximum of all models is 100% on each task. Audio Flamingo sets the new state-of-the-art results on most of these tasks.

trade-off between close-ended and open-ended tasks, limiting the overall performance (Deshmukh et al., 2023; Tang et al., 2023a). We use a different approach based on a widely adopted and stable method to train LLMs (Ouyang et al., 2022). Specifically, we use two training stages: pre-training and supervised fine-tuning (SFT), each with different subsets and training techniques. These innovations make Audio Flamingo achieve the state-of-the-art results on several audio understanding benchmarks with $< \frac{1}{3}$ number of parameters as Chu et al. (2023) and Gong et al. (2023c).

The third challenge is to give the audio language model the ability to quickly adapt to new tasks without fine-tuning, for instance, via in-context learning (ICL) (Brown et al., 2020) and retrieval. While recent audio language models have shown zero-shot abilities (Deshmukh et al., 2023; Gong et al., 2023c), they lack the ability to perform in-context few-shot learning to new tasks. In this paper, we introduce a series of techniques to realize this ability. We implement an efficient retrieval method, introduce an ICL template, and use retrieved samples to create interleaved ICL datasets for training. We also introduce a novel cross attention mask for interleaved samples. As a result, Audio Flamingo can be quickly adapted to new tasks via ICL and retrieval without task-specific fine-tuning. Our results confirm the efficacy of our approach and set new state-of-the-art few-shot benchmarks.

The last challenge is to give the audio language model the ability to chat with a user for many rounds. While prior methods have shown demos of dialogues (Gong et al., 2023c; Chu et al., 2023), they lack systematic and quantitative evidence. To address this challenge, we create two multi-turn dialogue datasets with GPT-4 (Achiam et al., 2023) based on detailed annotations of two datasets, with an emphasis on correlated context. We obtain a chat model by fine-tuning Audio Flamingo on these datasets. Our results show that our chat model has strong multi-turn dialogue ability and significantly outperforms previous methods.

We evaluate Audio Flamingo on a large and diverse set of close and open-ended benchmarks. A *single* Audio Flamingo model surpasses the previous state-of-the-art on most benchmarks, and the *chat* version of Audio Flamingo significantly outperforms baselines on dialogue benchmarks. Figure 1 summarizes the benchmark results of Audio Flamingo. We also briefly discuss about the neural architecture and hyper parameters in the experiments. Our key contributions include:

1. We propose Audio Flamingo: a Flamingo-based audio language model for audio understanding with a series of innovations. Audio Flamingo achieves state-of-the-art results on several close-ended and open-ended audio understanding tasks.
2. We design a series of methodologies for efficient use of ICL and retrieval, which lead to the state-of-the-art few-shot learning results.
3. We enable Audio Flamingo to have strong multi-turn dialogue ability, and show significantly better results compared to baseline methods.

2. Related work

Multimodal LLMs. There has been tremendous progress in the area of multimodal LLMs. In addition to text, these models take inputs from various modalities such as vision (Tsimpoukelli et al., 2021; Alayrac et al., 2022; Yang et al., 2023; Driess et al., 2023; Liu et al., 2023a; Li et al., 2023a), audio (Deshmukh et al., 2023; Gong et al., 2023b; Rubenstein et al., 2023), or multiple of them (Han et al., 2023; Tang et al., 2023b; Moon et al., 2023; Zhao et al., 2023), and each has a different integration method. In the audio modality, prior works have looked at speech tasks (Chen et al., 2023; Rubenstein et al., 2023), general audio understanding (Deshmukh et al., 2023; Gong et al., 2023c), music understanding (Gardner et al., 2023; Won et al., 2023; Li et al., 2023b; Liu et al., 2023b; Doh et al., 2023), or a combination of these (Gong et al., 2023b; Tang et al., 2023a; Chu et al., 2023). The focus of our paper is audio understanding, which includes non-speech sound and music, and non-verbal speech. Different from prior works, our model has stronger audio understanding ability, and is the first audio understanding model with

i) in-context few-shot learning ability, *ii)* retrieval augmented generation ability, and *iii)* strong multi-turn dialogue ability.

Audio encoders and representation. Many audio encoders extract audio features from the spectrogram, including CNN-based method (Kong et al., 2020) and Transformer-based methods (Gong et al., 2021; Chen et al., 2022; Défossez et al., 2022; Radford et al., 2023; Gong et al., 2023a). These methods are primarily targeted at solving a particular problem such as speech recognition or event detection. Based on these encoders, many joint audio-language embeddings have been proposed (Elizalde et al., 2023a;b; Wu et al., 2023; Huang et al., 2022; Li et al., 2023b). These methods use contrastive learning to map audio and language embeddings into the same space, and are often trained on a large variety of audio and language. However, many of these methods compute a single embedding for an audio and therefore may lose temporal information. In this paper, we build an audio encoder with sliding windows based on ClapCap (Elizalde et al., 2023b) to better capture long-range and temporal information.

Data augmentation. Due to limited amount of high-quality human annotated sounds besides speech transcriptions, many works have proposed to augment textural description with existing LLMs such as GPT-4 (Achiam et al., 2023). A common strategy is to provide an LLM with annotated tags, timestamps, and other miscellaneous information, and then ask it to generate captions (Wu et al., 2023; Doh et al., 2023; Mei et al., 2023; Gardner et al., 2023) or question-answering data pairs (Gong et al., 2023c;b; Liu et al., 2023b). In this paper, we leverage existing LLMs to generate two multi-turn dialogue datasets based on detailed annotations, which enable our model strong dialogue abilities.

In-context learning (ICL). In-context learning is a kind of few-shot learning ability, where an LLM rapidly adapts to a desired task at inference time only after looking at a few examples in the prompt (Brown et al., 2020). It has widely shown success in natural language tasks (Wei et al., 2021) and visual-language tasks (Alayrac et al., 2022; Yang et al., 2023). In the speech domain, ICL has been shown to help speech-related tasks such as speech recognition, translation, and processing (Gao et al., 2022; Wang et al., 2023; Hsu et al., 2023; Chen et al., 2023). However, ICL for general audio understanding is much less explored. In this paper, we propose the first audio understanding model with ICL ability.

Retrieval-augmented generation (RAG). Retrieval-augmented generation for LLMs is to improve generation quality by using external knowledge, for example from an external database, which contains useful and related knowledge. It has been widely applied in natural language tasks (Guu et al., 2020; Karpukhin et al., 2020; Lewis et al., 2020; Borgeaud et al., 2022) and visual-language models (Yang et al., 2023). In the audio-language domain, Ghosh et al. (2023) proposed a retrieval method for audio captioning by prepending captions from similar audios to the prompt. However, it does not provide the retrieved audio to the model. Consequently, the model loses information on how similar the retrieved audio is to the test audio. In contrast, we provide both the retrieved audio and text to our model. The benefit of this approach is that our model could determine when and how to use the retrieval based on the similarity between test and retrieved audio. We provide comparisons in our few-shot experiments.

3. Method

In this section, we introduce Audio Flamingo, an audio-understanding language model with few-shot learning via ICL and RAG. In Section 3.1, we introduce the architecture used in Audio Flamingo, including the audio feature extractor, audio representation transformation layers, language model, and the conditioning method. In Section 3.2, we introduce the training method of Audio Flamingo, including the training objective, design of masks, and training stages.

3.1. Architecture

Our neural architecture is composed of four components: *i)* an audio feature extractor with sliding window, *ii)* audio representation transformation layers, *iii)* a decoder-only language model, and *iv)* gated xattn-dense layers. Figure 2 summarizes the architecture.

***i)* Audio feature extractor with sliding window.** We use ClapCap (Elizalde et al., 2023b) as the audio feature extractor backbone, which we denote as $\mathcal{E}$. ClapCap is hard-coded to take 7-second of 44.1kHz raw audio as input, then transforms the audio into Mel-spectrogram of hop length 320, window length 1024, 64 Mel bins, and finally outputs a 1024-dimensional vector representation.

We consider each 7-second segment as a window and use sliding windows to extract features for longer audio. The overlap between consecutive windows is $7 \times 0.75 = 5.25$ seconds. Formally, let $x(s{:}t)$ be the segment of s to t seconds in audio x. Then, the extracted feature is $\left[\mathcal{E}(x(0{:}7)), \mathcal{E}(x(\frac{7}{4}{:}\frac{7\times5}{4})), \cdots, \mathcal{E}(x(\frac{7(m-1)}{4}{:}\frac{7(m+3)}{4})\right]$. The intuition of this design is to capture long-range and temporal information that might be ignored in a

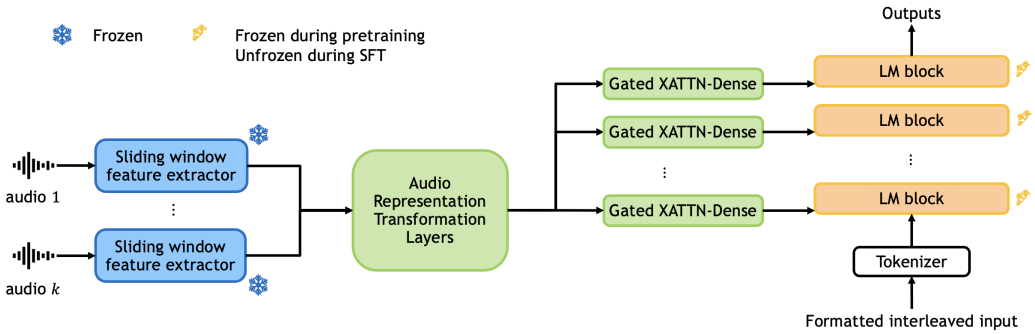


Figure 2. Neural architecture of Audio Flamingo. It takes interleaved audio and text as input and outputs free-form text.

single fused representation vector (Wu et al., 2023). We use a maximum of $m = 16$ sliding windows, which supports a maximum of 33.25 second audio length. [1] Long audio will be cropped and short audio will be zero-padded. If an entire segment is zero-padded then we will mask the corresponding embedding at cross attention. If the input is interleaved data with > 1 audio, we concatenate their sliding window representations.

***ii)* Audio representation transformation layers.** We increase model capacity by further applying a few audio representation transformation layers to the concatenated audio feature representations described earlier. It is comprised of 3 self-attention layers (Vaswani et al., 2017), with 8 heads and inner dimension 2048 each. This module is fully trainable.

***iii)* Language model.** We use a decoder-only causal LM in our architecture. In this paper, we use `OPT-IML-MAX-1.3B` (Iyer et al., 2022), a 1.3B parameter model with 24 LM blocks. It has been instruction-tuned on many natural language tasks.

***iv)* Conditioning LM on audio representations.** We use the gated xattn-dense layers from Flamingo (Alayrac et al., 2022) to achieve conditioning on audio inputs. Each layer has two blocks: 1) a residual block with cross attention and tanh gating, followed by 2) a residual block with dense layer and tanh gating. These layers are prepended to each LM block.

3.2. Training Method

***i)* Training objective.** Let x be the mono-channel audio input, y_{ins} be the instruction text (e.g. question), and y_{out} be the output text. For conciseness we use $z = (x, y_{\text{ins}}, y_{\text{out}})$ to represent each training sample.

We use maximum likelihood estimation (MLE) to train our model. Let $(y_{\text{out}})_t$ be the t-th token and $(y_{\text{out}})_{<t}$ the first $t-1$ tokens of the output. For a non-interleaved sample $z = (x, y_{\text{ins}}, y_{\text{out}})$, the log-likelihood is

$$\mathcal{L}(z) = \sum_{t=1}^{|y_{\text{out}}|} \log p_\theta \left((y_{\text{out}})_t | x, y_{\text{ins}}, (y_{\text{out}})_{<t}\right). \quad (1)$$

For an interleaved training sample composed of J samples $z_{\text{int}} = \{z^1, \cdots, z^J\}$, where $z^j = (x^j, y^j_{\text{ins}}, y^j_{\text{out}})$, the log-likelihood is computed over all outputs:

$$\mathcal{L}_{\text{int}}(z_{\text{int}} = \{z^1, \cdots, z^J\}) = \sum_{j=1}^{J} \sum_{t=1}^{|y^j_{\text{out}}|} \log p_\theta \left((y^j_{\text{out}})_t | z^{<j}, x^j, y^j_{\text{ins}}, (y^j_{\text{out}})_{<t}\right). \quad (2)$$

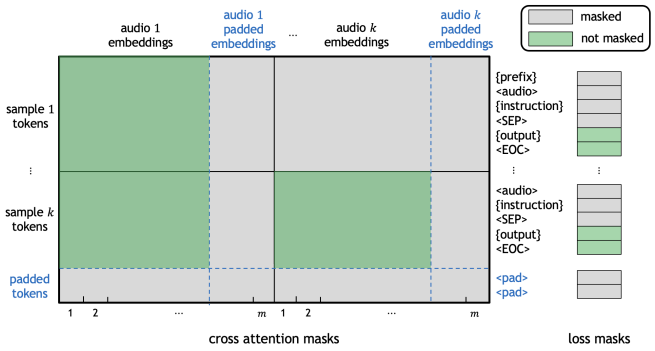
Note this interleaved data loss is different from prior models, which compute losses only on the last output y^J_{out} (Yang et al., 2023), or have either none or indirect conditioning on prior multimodal inputs $x^{<j}$ (Alayrac et al., 2022; Ghosh et al., 2023). We expect (2) can help the model look at a various number (including zero when $j = 1$) of ICL samples as well as the associated audio, thus improving robustness and training efficiency in a similar way as bucketing (Khomenko et al., 2016), especially when the ICL samples are retrieved similar samples. The corresponding loss mask is shown on the right-hand-side of Figure 3.

Let $\{\mathcal{D}^i, i \in \mathcal{I}\}$ be all non-interleaved training datasets, and $\{\mathcal{D}^{i'}_{\text{int}}, i' \in \mathcal{I}_{\text{int}}\}$ be all interleaved training datasets. The overall training objective is a weighted mixture of losses on each dataset:

$$L = -\sum_{i \in \mathcal{I}} \lambda_i \mathbb{E}_{z \sim \mathcal{D}^i} \mathcal{L}(z) - \sum_{i' \in \mathcal{I}_{\text{int}}} \lambda_{i'} \mathbb{E}_{z_{\text{int}} \sim \mathcal{D}^{i'}_{\text{int}}} \mathcal{L}_{\text{int}}(z_{\text{int}}), \quad (3)$$

where λ_i's are the weights for each dataset. The weights are constant hyper-parameters and have a huge impact on the final model. They are computed from the predefined number of epochs for each dataset (see Section 4.1 for the intuition, and Appendix A for details).

[1] We use $m = 16$ because most training samples are < 30s. Note that our cross attention complexity is linear in m and therefore the audio length. In prior self-attention methods the attention complexity is quadratic in the audio length.



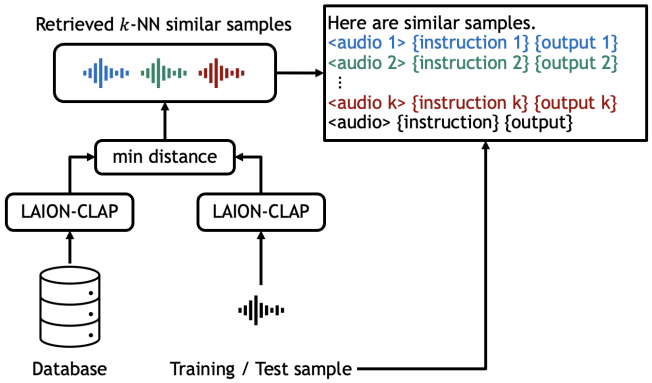


Figure 3. Left: the block upper-triangular cross attention masks between text tokens and audio embeddings. Right: the loss mask of an interleaved training sample.

Figure 4. Construction of ICL samples based on RAG. We use LAION-CLAP to find top-k most similar samples from the database, and use the retrieved audio and text to construct an ICL training sample.

***ii)* Cross attention masks.** We use block upper-triangular cross attention masks for interleaved samples so that the likelihood of j-th output $p_\theta(y^j_{\text{out}})$ is conditioned only on the first j audio inputs $x^{\leq j}$. We expect this helps making the model to look at previous audio. Figure 3 demonstrates the mask.

***iii)* Two training stages.** We divide training into pre-training and supervised fine-tuning (SFT), a widely adopted and stable method in training LMs (Ouyang et al., 2022). During pre-training we only train the audio representation transformation layers and the gated xattn-dense layers. The purpose is to obtain a good set of initialization weights for these layers. During SFT, we unfreeze the entire LM, and train all modules but the audio encoder. [2]

4. Data

4.1. Datasets

In this section, we introduce our data strategies, including dataset collection, generation, and blending. We also introduce templates for each type of dataset.

Dataset sources. We train our model on a variety of audio datasets that can be roughly classified into three types: music, non-speech general sound, and non-verbal speech. In this paper, we focus on these data given the immediate availability of state-of-the-art speech transcription models. We look at three types of tasks: (1) *audio captioning* (CAP), where we would like the model to describe the audio in a sentence; (2) *audio question-answering* (AQA), where we would like the model to answer questions regarding the audio, and (3) *audio classification* (CLS), where we would like the model to classify the sound into one or more labels corresponding to events, scenes, music genres, instruments, qualities, and others. An overview of all training datasets is shown in Table 1.

ICL datasets. In order to give our model in-context learning and retrieval augmentation abilities, we construct ICL datasets for each of the raw datasets based on kNN computed on audio embeddings. Let $\mathcal{D}^i$ be the i-th training dataset. For each $z = (x, y_{\text{ins}}, y_{\text{out}}) \in \mathcal{D}^i$, we find its top-$k$ closest training samples in $\mathcal{D}^i$ excluding z, where the distance function is ℓ_2 in the fused LAION-CLAP embedding space (Wu et al., 2023) for the audio part of the sample. We use Faiss-gpu (Johnson et al., 2019) to accelerate searching. Figure 4 demonstrates this process.

Dataset staging and blending. We use different datasets during the pre-training stage and the supervised fine-tuning (SFT) stage. The selection is based on data quality, diversity, source, and size as described below. 1) Data quality: low quality datasets, including those with low-quality or noisy audio, low-quality text, and inaccurate text annotation, are used for pre-training. 2) Data diversity: datasets with less diversity or strong biases in label distributions are used for pre-training. 3) Data sources: datasets containing AI-generated contents are mostly used for pre-training, whereas some high-quality subsets may be used for SFT. 4) Data sizes: very large datasets may be used both for pre-training and SFT. 5) ICL datasets are used in the SFT stage.

We assign different weights λ_i when sampling from different datasets based on their sizes, quality, and diversity. The weights are computed from the number of epochs for each dataset. The details of staging and weights can be found in Appendix A.

[2] In initial experiments we found unfreezing the audio encoder caused training instability.

Table 1. All datasets used to train our model. The total number of audio-text pairs is approximately 5.9 million. The total length of audio is approximately 18.1 thousand hours.

Audio Type	Task	Datasets	#Audio-Text Pairs
General Sound	CAP	WavCaps (Mei et al., 2023), Macs (Martin Morato & Mesaros, 2021), SoundDescs (Oncescu et al., 2021), Clotho-v2 (Drossos et al., 2020), WavText5K (Deshmukh et al., 2022), LAION-630k (Wu et al., 2023)	∼829 K
	AQA	Clotho-AQA (Lipping et al., 2022), Open-AQA (Gong et al., 2023b)	∼1970 K
	CLS	AudioSet (Gemmeke et al., 2017), FSD50k (Fonseca et al., 2021), CochlScene (Jeong & Park, 2022), NonSpeech7K (Rashid et al., 2023), Chime-Home (Foster et al., 2015), Sonyc-UST (Cartwright et al., 2019)	∼1091 K
Music	CAP	LP-MusicCaps (Doh et al., 2023), MusicCaps (Agostinelli et al., 2023)	∼1389 K
	AQA	MusicQA (Liu et al., 2023b), MusicAVQA (Li et al., 2022)	∼94 K
	CLS	NSynth (Engel et al., 2017), MTG-Jamendo (Bogdanov et al., 2019), FMA (Defferrard et al., 2016), MusDB-HQ (Rafii et al., 2019),	∼459 K
Speech	CLS	MSP-Podcast (Lotfian & Busso, 2017), Emov-DB (Adigwe et al., 2018) JL-Corpus (James et al., 2018), Tess (Pichora-Fuller & Dupuis, 2020), MELD (Poria et al., 2018), OMGEmotion (Barros et al., 2018)	∼92 K

4.2. Templates

Our templates are based on OPT-IML's template (Iyer et al., 2022) and Flamingo's multimodal template (Alayrac et al., 2022). For a non-interleaved sample, the template is describe below.

```
<s>{task description}<audio>{instruction}
Options:\n- option_1\n··· - option_m
<SEP>{output}<EOC></s>
```

In this template, `<audio>` is the special token that informs the language model the location of audio in the context. The `{task description}` is natural language that tells the language model which task it is handling, for example "*The task is event classification*". The `{instruction}` is the language instruction such as a *question*. The `options` sentence is to tell the language model all options for classification so that it can classify an audio by outputting free-form text. The `{output}` is the ground truth output that will be trained. The `<SEP>` token is a separator that indicates the end of instruction, and `<EOC>` is the end-of-chunk token that indicates the end of a sample. Below is the template for interleaved (ICL) samples with $k+1$ tuples of (audio, instruction, output).

```
<s>{task description}Here are similar samples.
<audio>{instruction_1}<SEP>{output_1}<EOC>
···
<audio>{instruction_k}<SEP>{output_k}<EOC>
<audio>{instruction}
Options:\n- option_1\n··· - option_m
<SEP>{output}<EOC></s>
```

4.3. Multi-Turn Dialogue Dataset

We aim at giving our model stronger abilities in dealing with complicated multi-turn dialogues. To achieve this, we use GPT-4 (Achiam et al., 2023) to generate two multi-turn dialogue datasets. We construct these datasets based on the strongly labeled AudioSet-SL (Hershey et al., 2021) and MusicCaps (Agostinelli et al., 2023), with thresholding based on LAION-CLAP embeddings (Wu et al., 2023) to filter undesirable samples. We name these two generated datasets `AF-Dialogue-AudioSetSL` and `AF-Dialogue-MusicCaps`, respectively. The detailed instructions, filtering method, dataset statistics, and examples are in Appendix B. We use the following template for an s-turn dialogue data sample.

```
<s>The task is dialogue.<audio>
user: {instruction_1}
assistant: <SEP>{output_1}<EOC>
···
user: {instruction_s}
assistant: <SEP>{output_s}<EOC></s>
```

5. Experiments

In this section, we answer the following questions:
Q1. Does Audio Flamingo understand audio better than the state-of-the-art baselines?
Q2. How significantly does the ICL-based RAG help Audio Flamingo adapt to new tasks?
Q3. What is Audio Flamingo's ability to have multi-turn dialogues with a user?
Q4. Which specific configuration of Audio Flamingo works the best overall?

5.1. Experimental Setup

We use 8 NVIDIA A100 GPUs to train our model. During pre-training, we use batch size = 384, AdamW optimizer (Loshchilov & Hutter, 2017) with learning rate = 1×10^{-4} and weight decay = 0.1. For efficiency, we use *bf16* with automatic mixed precision. During

supervised fine-tuning (SFT), we reduce the batch size to 128, the learning rate to 2×10^{-5}, and use *fp32* for better numerical precision. We let the maximum number of interleaved samples to be 8 unless specified. We set the maximum number of text tokens to be 512.

We compare to the most recent state-of-the-art baselines on several benchmarks. On each dataset, we choose the best score among all SOTA baselines as the reference value. Unless specified, we report accuracy for question-answering and single-label classification, F1 for multi-label classification, and CIDEr (Vedantam et al., 2015) for captioning and dialogues. Note we use free-form text output to evaluate our model at all times. We use a *single* model to evaluate on all benchmarks except for dialogues, and a *chat* model on dialogues.

For zero-shot and few-shot benchmarks, these datasets are excluded from the pre-training sets and SFT sets. For those derived from a parent dataset (e.g. AudioCaps audio are derived from AudioSet), we removed the training samples from the parent set as well as other child sets derived from that parent set.

5.2. Q1: Strong Audio Understanding Ability

We evaluate our model on several in-distribution (train-test) benchmarks, and compare to state-of-the-art audio language model baselines. The results are shown in Table 2. Note that we define $\text{F1}_{\text{approx}}$ to measure inexact but similar predicted labels in FSD50k, where we consider the prediction to be correct if the sentence BERT similarity between output and ground truth is > 0.8 (Reimers & Gurevych, 2019; 2020). This metric is applied to outputs from baselines as well.

Audio Flamingo can match or outperform SOTA baselines – many of which are much larger LLMs (7B (Gong et al., 2023c;b; Chu et al., 2023) or 13B (Tang et al., 2023a)) – on most tasks, indicating our proposed method has strong audio understanding ability. Our model also *listens to the audio* better. On ClothoAQA, our model has 10.4% higher accuracy than baselines on numerical question answering, indicating our model understands the number of occurrences better. On NSynth, our model has 20.4% higher F1 on quality prediction and 18.6% higher accuracy on source prediction, indicating our model understands the overall quality of audio better. In Appendix C.3, we use qualitative samples to show that our model understands the order of appearance of different sounds, perceives loudness and its change over time, and perceives the distances of sounds from different objects.

5.3. Q2: In-Context Few-Shot Learning

We aim to measure the effect of ICL-based RAG in Audio Flamingo when it is evaluated on unseen datasets.

First, we report the results on several zero-shot benchmarks and comparison with SOTA zero-shot methods in Table 3. The results indicate our method is better on most tasks and has strong generalization ability.

We then apply ICL-based RAG to these benchmarks. We compare to our zero-shot results and the SOTA baseline of audio captioning on AudioCaps. The results on classification are shown in Table 4, and the comparison on retrieval-augmented audio captioning is shown in Table 5. As expected, there is consistent improvement over zero-shot results, with an average improvement over 10% for classification. Our method also significantly outperforms the SOTA retrieval-augmented audio captioning method on AudioCaps. In Appendix C.1, we show Audio Flamingo can adapt to unseen labels. In Appendix C.4, we show Audio Flamingo *looks at related retrieval* (e.g., by taking key words from retrieved captions), and *ignores noisy retrieval*.

5.4. Q3: Multi-Turn Dialogues

We measure Audio Flamingo's ability to answer questions in a multi-turn dialogue setting. The context is more complex and strongly correlated between rounds (e.g. there exist many pronouns and follow-up questions). We fine-tune Audio Flamingo on the two sets that we generated (`AF-Dialogue-AudioSetSL` and `AF-Dialogue-MusicCaps`) to obtain a chat model. We evaluate the chat model on the test split of these two dialogue datasets. We take user instructions and let the model generate answers turn-by-turn (where previous generated answers become the chatting history for next generation). We compare to Qwen-Audio (Chu et al., 2023), LTU (Gong et al., 2023c), and MU-LLaMA (Liu et al., 2023b) in Table 6. [3] Our chat model achieves significantly better results than baseline methods. In Appendix C.5, we use qualitative samples to show that our chat model *captures context* such as prior information and pronouns better.

5.5. Q4: Ablation Studies

Effect of number of few-shot samples. We study different numbers of in-context few-shot samples and evaluate on the few-shot benchmarks. In Figure 5, we plot the relative improvements over zero-shot results. Results show a clear trend that having more ICL samples improves few-shot results, and the improvement

[3] While the baseline methods claimed to support multi-turn dialogues, we were unable to find quantitative evidence.

Table 2. Evaluation of Audio Flamingo versus SOTA baseline methods on in-distribution benchmarks. Reference values are the SOTA models for each task. Audio Flamingo shows strong audio understanding ability on captioning, question answering, and audio classification.

Dataset	Task	Metric	Previous SOTA ↑	Ours ↑
Clotho-v2	CAP	CIDEr	0.441 (Chu et al., 2023)	**0.465**
$\text{ClothoAQA}_{\text{unanimous}}$	AQA	ACC	74.9% (Chu et al., 2023)	**86.9**%
$\text{ClothoAQA}_{\text{non-binary}}$	AQA	ACC	29.1% (Deshmukh et al., 2023)	**49.5**%
$\text{ClothoAQA}_{\text{numerical}}$	AQA	ACC	26.2% (Deshmukh et al., 2023)	**36.4**%
$\text{MusicAVQA}_{\text{audio-only}}$	AQA	ACC	**72.1**% (Chu et al., 2023)	71.6%
CochlScene	CLS	ACC	**91.6**% (Deshmukh et al., 2023)	83.0%
NonSpeech7k	CLS	ACC	79.0% (Rashid et al., 2023)	**85.1**%
FSD50k	CLS	$\text{F1}_{\text{approx}}$	65.6% (Deshmukh et al., 2023)	**69.7**%
$\text{NS}_{\text{instrument}}$	CLS	ACC	**78.8**% (Chu et al., 2023)	77.1%
$\text{NS}_{\text{quality}}$	CLS	F1	46.3% (Deshmukh et al., 2023)	**66.7**%
$\text{NS}_{\text{source}}$	CLS	ACC	60.1% (Deshmukh et al., 2023)	**78.7**%

Table 3. Evaluation of Audio Flamingo versus SOTA baseline methods on zero-shot benchmarks. Reference values are the SOTA models for each task. Audio Flamingo shows strong zero-shot generalization ability.

Dataset	Task	Metric	Previous SOTA (0-shot) ↑	Ours (0-shot) ↑
AudioCaps (Kim et al., 2019)	CAP	CIDEr	0.281 (Salewski et al., 2023)	**0.502**
CREMA-D (Cao et al., 2014)	CLS	ACC	18.5% (Deshmukh et al., 2023)	**26.5**%
Ravdess (Livingstone & Russo, 2018)	CLS	ACC	**21.7**% (Elizalde et al., 2023b)	20.9%
US8K (Salamon et al., 2014)	CLS	ACC	71.9% (Deshmukh et al., 2023)	**75.0**%
GTZAN (Sturm, 2013)	CLS	ACC	**71.0**% (Han et al., 2023)	67.9%
Medley-solos-DB (Lostanlen et al., 2019)	CLS	ACC	61.3% (Deshmukh et al., 2023)	**92.7**%

Table 4. Evaluation of few-shot results of Audio Flamingo with ICL-based RAG. Δ is the *absolute* improvement of few-shot over zero-shot results in Table 3. ICL-based RAG leads to consistent improvement over zero-shot results.

Dataset	Ours (8-shot) ↑	Δ ↑
CREMA-D	31.8%	5.3%
Ravdess	35.2%	14.3%
US8K	94.7%	19.4%
GTZAN	79.5%	11.6%
Medley-solos-DB	95.7%	3.0%
Dataset	**Ours (16-shot) ↑**	**Δ ↑**
CREMA-D	35.4%	8.9%
Ravdess	40.2%	19.3%
US8K	91.9%	16.9%
GTZAN	76.3%	8.4%
Medley-solos-DB	96.0%	3.3%

highly depends on the dataset.

Effect of architecture. In Figure 6, we compare results between `opt-1.3b` (Zhang et al., 2022) and the instruction-tuned `opt-iml-max-1.3b` (Iyer et al., 2022), and also compare different audio encoders including ClapCap (Elizalde et al., 2023b), Clap2023 (Elizalde et al., 2023b), and LAION-CLAP (Wu et al., 2023). In terms of the LM backbone, the instruction-tuned `opt` is better in most tasks. As of audio encoders, LAION-CLAP is worse in most tasks, ClapCap is better in open-ended tasks, and Clap2023 is better in most close-ended tasks.

Table 5. Evaluation of retrieval-augmented audio captioning on AudioCaps. We compare Audio Flamingo to the SOTA baseline RECAP (Ghosh et al., 2023). Audio Flamingo achieves significantly better results than RECAP.

Method	RECAP	Ours	Ours	Ours
# Shots	4	4	8	16
CIDEr ↑	0.359	0.518	0.538	**0.546**

Effect of training data. In Figure 7, we compare Audio Flamingo trained on three different training sets to the Pengi baseline: (1) no pretraining, SFT on our best available dataset that is strictly a subset of Pengi's training set, (2) pretraining and SFT on this strict subset, and (3) pretraining and SFT on our curated datasets. Audio Flamingo achieves better evaluation results than Pengi even if no extra data is used. In addition, increasing the data amount in pretraining and SFT can improve the results on average.

Table 6. Evaluation of Audio Flamingo versus baseline methods on the multi-turn dialogue test sets. `A` stands for `AF-Dialogue-AudioSetSL`, `M` stands for `AF-Dialogue-MusicCaps`, and the superscript H stands for an additional held-out testset generated with `gpt-3.5-turbo`. We report CIDEr, Bleu4 (Papineni et al., 2002), and Rouge-L (R-L) (Lin, 2004). Methods with the † superscript are fine-tuned on our dialogue training sets, and methods without † are evaluated zero-shot on the dialogue test sets. Audio Flamingo significantly outperforms larger baseline models in all settings, indicating strong dialogue ability of our proposed model.

Testset	Method	CIDEr ↑	Bleu4 ↑	R-L ↑
`A`	Qwen-Audio	0.507	0.060	0.292
`A`	LTU	0.580	0.122	0.324
`A`	LTU†	0.823	0.153	0.403
`A`	Ours†	**1.622**	**0.237**	**0.473**
$\texttt{A}^{\texttt{H}}$	LTU†	0.523	0.095	0.343
$\texttt{A}^{\texttt{H}}$	Ours†	**1.904**	**0.219**	**0.476**
`M`	MU-LLaMA	0.585	0.083	0.348
`M`	LTU	0.168	0.065	0.217
`M`	LTU†	0.419	0.108	0.336
`M`	Ours†	**1.143**	**0.142**	**0.417**
$\texttt{M}^{\texttt{H}}$	LTU†	0.558	0.083	0.347
$\texttt{M}^{\texttt{H}}$	Ours†	**1.350**	**0.207**	**0.448**

6. Conclusion and Future Work

In this paper, we present Audio Flamingo, an audio language model with a series of innovations that achieves the state-of-the-art results on several close-ended and open-ended audio understanding tasks without task specific fine-tuning. It also has strong ICL and RAG abilities, and has the state-of-the-art few-shot learning results. Furthermore, we design a dataset generation strategy and introduce two dialogue datasets, enabling Audio Flamingo to chat with a user about the audio for multiple rounds and achieve state-of-the-art results on dialogue benchmarks.

Our model has several limitations, which we plan to address in future work. One important future direction is to investigate scaling strategies for using larger LMs. Assuming that larger LMs could have better knowledge and stronger ability to follow instructions, we believe that Audio Flamingo could benefit from a larger LM. A second future direction is to investigate complex speech-related tasks beyond transcription. This requires our model to condition on dense embeddings (Chen et al., 2023). This modification is straightforward in Audio Flamingo as the architecture is flexible enough to support new embeddings through the addition of new cross-attention heads. Another future direction is to build a audio language model that can output both text and audio and follow more complex interleaved instructions. Finally, a future direction towards unifying more modalities is to combine the audio understanding abilities of our model with visual language models (Alayrac et al., 2022) such that one model could understand image, video, and audio.

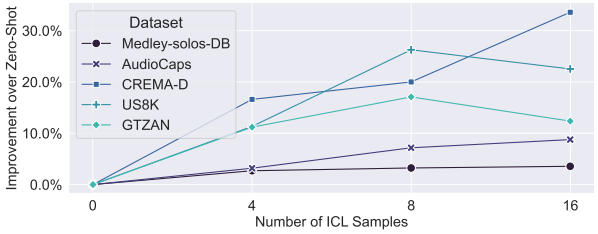


Figure 5. Relative improvement of few-shot results over zero-shot results under different number of ICL samples. Using more ICL samples consistently improves few-shot results, and the benefit is dataset-dependent.

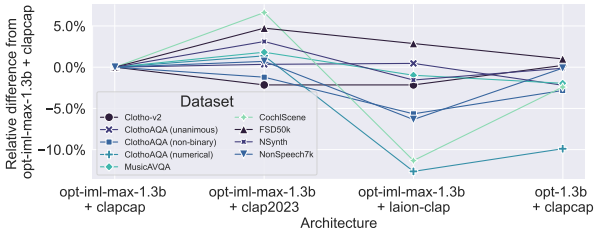


Figure 6. Relative difference of results from different LM backbones and audio encoders. The `opt-1.3b` model without instruction-tuning is systematically worse than the instruction-tuned `opt-iml-max-1.3b`. LAION-CLAP is worse than ClapCap in most cases. Clap2023 is better than ClapCap in close-ended tasks, but worse in open-ended tasks including captioning and open question-answering.

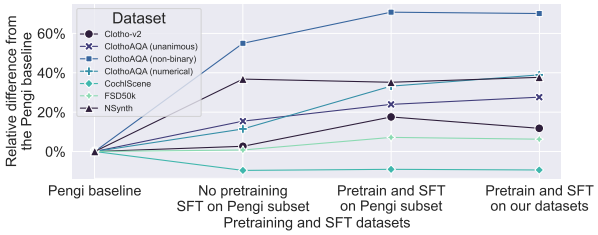


Figure 7. Relative difference of results from different training datasets compared to the Pengi baseline. Audio Flamingo achieves better results than Pengi even if no extra data is used. Larger scale pretraining and SFT lead to better results on these benchmarks on average.

Acknowledgement

We thank Siddharth Gururani, Zihan Liu, Mostofa Patwary, Shrimai Prabhumoye, and Chen Zhu for helpful discussions. We thank Ke Chen and Yuan Gong for help on sharing datasets with us.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. The proposed method facilitates automation in the audio-language domain and may be used in a variety of scenarios such as education, healthcare, environment, industry, music, etc. Careful use of the model is essential to ensure compliance with copyright restrictions in certain situations.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Adigwe, A., Tits, N., Haddad, K. E., Ostadabbas, S., and Dutoit, T. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*, 2018.

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 preprint arXiv:2301.11325*, 2023.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Barros, P., Churamani, N., Lakomkin, E., Siqueira, H., Sutherland, A., and Wermter, S. The omg-emotion behavior dataset. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2018.

Bogdanov, D., Won, M., Tovstogan, P., Porter, A., and Serra, X. The mtg-jamendo dataset for automatic music tagging. ICML, 2019.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

Cartwright, M., Mendez, A. E. M., Cramer, A., Lostanlen, V., Dove, G., Wu, H.-H., Salamon, J., Nov, O., and Bello, J. Sonyc urban sound tagging (sonyc-ust): A multilabel dataset from an urban acoustic sensor network. 2019.

Chen, K., Du, X., Zhu, B., Ma, Z., Berg-Kirkpatrick, T., and Dubnov, S. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 646–650. IEEE, 2022.

Chen, Z., Huang, H., Andrusenko, A., Hrinchuk, O., Puvvada, K. C., Li, J., Ghosh, S., Balam, J., and Ginsburg, B. Salm: Speech-augmented language model with in-context learning for speech recognition and translation. *arXiv preprint arXiv:2310.09424*, 2023.

Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., Yan, Z., Zhou, C., and Zhou, J. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*, 2023.

Defferrard, M., Benzi, K., Vandergheynst, P., and Bresson, X. Fma: A dataset for music analysis. *arXiv preprint arXiv:1612.01840*, 2016.

Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Deshmukh, S., Elizalde, B., and Wang, H. Audio retrieval with wavtext5k and clap training. *arXiv preprint arXiv:2209.14275*, 2022.

Deshmukh, S., Elizalde, B., Singh, R., and Wang, H. Pengi: An audio language model for audio tasks. *arXiv preprint arXiv:2305.11834*, 2023.

Doh, S., Choi, K., Lee, J., and Nam, J. Lp-musiccaps: Llm-based pseudo music captioning. *arXiv preprint arXiv:2307.16372*, 2023.

Driess, D., Xia, F., Sajjadi, M. S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

Drossos, K., Lipping, S., and Virtanen, T. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740. IEEE, 2020.

Duan, H., Wei, J., Wang, C., Liu, H., Fang, Y., Zhang, S., Lin, D., and Chen, K. Botchat: Evaluating llms' capabilities of having multi-turn dialogues. *arXiv preprint arXiv:2310.13650*, 2023.

Elizalde, B., Deshmukh, S., Al Ismail, M., and Wang, H. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023a.

Elizalde, B., Deshmukh, S., and Wang, H. Natural language supervision for general-purpose audio representations, 2023b. URL https://arxiv.org/abs/2309.05767.

Engel, J., Resnick, C., Roberts, A., Dieleman, S., Eck, D., Simonyan, K., and Norouzi, M. Neural audio synthesis of musical notes with wavenet autoencoders, 2017.

Fonseca, E., Favory, X., Pons, J., Font, F., and Serra, X. Fsd50k: an open dataset of human-labeled sound events. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852, 2021.

Foster, P., Sigtia, S., Krstulovic, S., Barker, J., and Plumbley, M. D. Chime-home: A dataset for sound source recognition in a domestic environment. In *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5. IEEE, 2015.

Gao, H., Ni, J., Qian, K., Zhang, Y., Chang, S., and Hasegawa-Johnson, M. Wavprompt: Towards few-shot spoken language understanding with frozen language models. *arXiv preprint arXiv:2203.15863*, 2022.

Gardner, J., Durand, S., Stoller, D., and Bittner, R. M. Llark: A multimodal foundation model for music. *arXiv preprint arXiv:2310.07160*, 2023.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Ghosh, S., Kumar, S., Evuru, C. K. R., Duraiswami, R., and Manocha, D. Recap: Retrieval-augmented audio captioning. *arXiv preprint arXiv:2309.09836*, 2023.

Gong, Y., Chung, Y.-A., and Glass, J. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

Gong, Y., Khurana, S., Karlinsky, L., and Glass, J. Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers. *arXiv preprint arXiv:2307.03183*, 2023a.

Gong, Y., Liu, A., Luo, H., Karlinsky, L., and Glass, J. Joint audio and speech understanding. In *IEEE Automatic Speech Recognition and Understanding Workshop*, 2023b.

Gong, Y., Luo, H., Liu, A. H., Karlinsky, L., and Glass, J. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*, 2023c.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Han, J., Zhang, R., Shao, W., Gao, P., Xu, P., Xiao, H., Zhang, K., Liu, C., Wen, S., Guo, Z., et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023.

Hershey, S., Ellis, D. P., Fonseca, E., Jansen, A., Liu, C., Moore, R. C., and Plakal, M. The benefit of temporally-strong labels in audio event classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 366–370. IEEE, 2021.

Hsu, M.-H., Chang, K.-W., Li, S.-W., and Lee, H.-y. An exploration of in-context learning for speech language model. *arXiv preprint arXiv:2310.12477*, 2023.

Huang, Q., Jansen, A., Lee, J., Ganti, R., Li, J. Y., and Ellis, D. P. Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415*, 2022.

Huang, R., Li, M., Yang, D., Shi, J., Chang, X., Ye, Z., Wu, Y., Hong, Z., Huang, J., Liu, J., et al. Audiogpt: Understanding and generating speech, music, sound, and talking head. *arXiv preprint arXiv:2304.12995*, 2023.

Iyer, S., Lin, X. V., Pasunuru, R., Mihaylov, T., Simig, D., Yu, P., Shuster, K., Wang, T., Liu, Q., Koura, P. S., et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022.

James, J., Tian, L., and Watson, C. An open source emotional speech corpus for human robot interaction applications. *Interspeech 2018*, 2018.

Jeong, I.-Y. and Park, J. Cochlscene: Acquisition of acoustic scene data using crowdsourcing. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 17–21. IEEE, 2022.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.

Khomenko, V., Shyshkov, O., Radyvonenko, O., and Bokhan, K. Accelerating recurrent neural network training using sequence bucketing and multi-gpu data parallelization. In *2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP)*, pp. 100–103. IEEE, 2016.

Kim, C. D., Kim, B., Lee, H., and Kim, G. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, 2019.

Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., and Plumbley, M. D. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Li, G., Wei, Y., Tian, Y., Xu, C., Wen, J.-R., and Hu, D. Learning to answer questions in dynamic audio-visual scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19108–19118, 2022.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023a.

Li, Y., Yuan, R., Zhang, G., Ma, Y., Chen, X., Yin, H., Lin, C., Ragni, A., Benetos, E., Gyenge, N., et al. Mert: Acoustic music understanding model with large-scale self-supervised training. *arXiv preprint arXiv:2306.00107*, 2023b.

Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Lipping, S., Sudarsanam, P., Drossos, K., and Virtanen, T. Clotho-aqa: A crowdsourced dataset for audio question answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pp. 1140–1144. IEEE, 2022.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023a.

Liu, S., Hussain, A. S., Sun, C., and Shan, Y. Music understanding llama: Advancing text-to-music generation with question answering and captioning. *arXiv preprint arXiv:2308.11276*, 2023b.

Livingstone, S. R. and Russo, F. A. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lostanlen, V., Cella, C.-E., Bittner, R., and Essid, S. Medley-solos-DB: a cross-collection dataset for musical instrument recognition, February 2019. URL https://doi.org/10.5281/zenodo.1344103.

Lotfian, R. and Busso, C. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, 2017.

Lyu, C., Wu, M., Wang, L., Huang, X., Liu, B., Du, Z., Shi, S., and Tu, Z. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*, 2023.

Martin Morato, I. and Mesaros, A. Diversity and bias in audio captioning datasets. 2021.

Mei, X., Meng, C., Liu, H., Kong, Q., Ko, T., Zhao, C., Plumbley, M. D., Zou, Y., and Wang, W. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv preprint arXiv:2303.17395*, 2023.

Mohanty, S. P. Sound Of 114 Species Of Birds Till 2022. URL https://www.kaggle.com/datasets/soumendraprasad/sound-of-114-species-of-birds-till-2022.

Moon, S., Madotto, A., Lin, Z., Nagarajan, T., Smith, M., Jain, S., Yeh, C.-F., Murugesan, P., Heidari, P., Liu, Y., et al. Anymal: An efficient and scalable any-modality augmented language model. *arXiv preprint arXiv:2309.16058*, 2023.

Oncescu, A.-M., Koepke, A., Henriques, J. F., Akata, Z., and Albanie, S. Audio retrieval with natural language queries. *arXiv preprint arXiv:2105.02192*, 2021.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Park, J., Cho, Y., Sim, G., Lee, H., and Choo, J. Enemy spotted: In-game gun sound dataset for gunshot classification and localization. In *2022 IEEE Conference on Games (CoG)*, pp. 56–63. IEEE, 2022.

Pichora-Fuller, M. K. and Dupuis, K. Toronto emotional speech set (TESS), 2020. URL https://doi.org/10.5683/SP2/E8H2MF.

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., and Mihalcea, R. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023.

Rafii, Z., Liutkus, A., Stöter, F.-R., Mimilakis, S. I., and Bittner, R. Musdb18-hq - an uncompressed version of musdb18, August 2019. URL https://doi.org/10.5281/zenodo.3338373.

Rashid, M. M., Li, G., and Du, C. Nonspeech7k dataset: Classification and analysis of human non-speech sound. *IET Signal Processing*, 17(6):e12233, 2023.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Reimers, N. and Gurevych, I. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2020. URL https://arxiv.org/abs/2004.09813.

Rubenstein, P. K., Asawaroengchai, C., Nguyen, D. D., Bapna, A., Borsos, Z., Quitry, F. d. C., Chen, P., Badawy, D. E., Han, W., Kharitonov, E., et al. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*, 2023.

Salamon, J., Jacoby, C., and Bello, J. P. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 1041–1044, 2014.

Salewski, L., Fauth, S., Koepke, A., and Akata, Z. Zero-shot audio captioning with audio-language model guidance and audio context keywords. *arXiv preprint arXiv:2311.08396*, 2023.

Sturm, B. L. The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461*, 2013.

Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., and Zhang, C. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023a.

Tang, Z., Yang, Z., Khademi, M., Liu, Y., Zhu, C., and Bansal, M. Codi-2: In-context, interleaved, and interactive any-to-any generation. *arXiv preprint arXiv:2311.18775*, 2023b.

Tsimpoukelli, M., Menick, J. L., Cabi, S., Eslami, S., Vinyals, O., and Hill, F. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Wang, S., Yang, C.-H. H., Wu, J., and Zhang, C. Can whisper perform speech-based in-context learning. *arXiv preprint arXiv:2309.07081*, 2023.

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., and Le, Q. V. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Won, M., Hung, Y.-N., and Le, D. A foundation model for music informatics. *arXiv preprint arXiv:2311.03318*, 2023.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Yang, Z., Ping, W., Liu, Z., Korthikanti, V., Nie, W., Huang, D.-A., Fan, L., Yu, Z., Lan, S., Li, B., et al. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. In *EMNLP*, 2023.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Zhao, Z., Guo, L., Yue, T., Chen, S., Shao, S., Zhu, X., Yuan, Z., and Liu, J. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv preprint arXiv:2305.16103*, 2023.

A. Dataset Staging, Weights, and Templates

Table 1 includes an overview of datasets (by type) we use to train Audio Flamingo. We construct instructions for each task and dataset. Below are all instruction templates we use.

Audio Captioning:
- `Describe the sound/music in a sentence.`
- `Describe the sound/music at length.`

Audio Question Answering:
- `{question}`
- `Please answer this question: {question}`
- `Please answer this question: {question}. Options:\n- yes\n- no`

Audio Classification:
- `Classify this sound. (Options: ...)`
- `Describe the sound in {number} words.`
- `What is the emotion of this speech? (Options: ...)`
- `What is the instrument/genre of this music? (Options: ...)`
- `This music note is produced by`

The detailed pre-training datasets and their number of epochs are shown in Table 7. The detailed SFT datasets and their number of epochs are shown in Table 8.

Table 7. Pre-training datasets and epochs.

Dataset	Audio Length	#Audio-Text Pairs	Epochs
OpenAQA	693.2 hrs	1959.8K	1.0
$\text{Laion630k}_{\text{BBCSoundEffects}}$	456.9 hrs	15.1K	5.0
$\text{Laion630k}_{\text{Freesound}}$	2494.8 hrs	306.5K	1.0
SoundDescs	749.7 hrs	23.1K	1.0
WavCaps	3793.3 hrs	402.6 K	1.75
AudioSet	2617.8 hrs	950.8K	1.0
WavText5K	23.8 hrs	4.3K	3.0
MSP-Podcast	73.9 hrs	45.1K	1.2
MELD	8.7 hrs	32.9K	2.4
$\text{MusicAVQA}_{\text{audio-visual}}$	142.4 hrs	17.9K	3.0
MusicQA	62.9 hrs	70K	1.2
$\text{LP-MusicCaps}_{\text{MSD}}$	5805.7 hrs	1331.8K	1.0
NSynth	321.3 hrs	289.2K	0.4
MTG-Jamendo	3768.9 hrs	55.6K	1.0

Table 8. SFT datasets and epochs.

Dataset	Audio Length	#Audio-Text Pairs	Epochs	ICL Dataset Epochs
ClothoAQA	7.4 hrs	9.7K	3.5	0.5
OpenAQA	693.2 hrs	1959.8K	0.1	-
Clotho-v2	24.0 hrs	19.2K	2.0	0.5
$\text{Laion630k}_{\text{Epidemic}}$	209.4 hrs	40.7K	0.8	0.2
MACS	10.9 hrs	17.3K	0.8	0.2
FSD50k	80.8 hrs	41.0K	0.9	0.3
CochlScene	169.0 hrs	60.9K	1.2	0.3
NonSpeech 7k	6.2 hrs	6.3K	2.4	0.6
Chime-home	5.0 hrs	4.5K	1.5	0.5
Sonyc-UST	34.9 hrs	27.9K	0.8	0.2
Emov-DB	7.8 hrs	6.8K	1.6	0.4
JL-Corpus	1.4 hrs	2.4K	6.0	1.5
Tess	1.6 hrs	2.8K	2.0	0.5
OMGEmotion	3.0 hrs	1.7K	3.0	-
$\text{MusicAVQA}_{\text{audio-only}}$	77.1 hrs	5.7K	5.0	1.0
MusicQA	62.9 hrs	70K	0.35	0.05
$\text{LP-MusicCaps}_{\text{MSD}}$	5805.7 hrs	1331.8K	0.025	0.007
$\text{LP-MusicCaps}_{\text{MTT}}$	126.4 hrs	46.9K	0.8	0.2
$\text{LP-MusicCaps}_{\text{MC}}$	7.4 hrs	7.9K	2.0	-
MusicCaps	7.4 hrs	2.6K	6.0	-
NSynth	321.3 hrs	289.2K	1.0	1.0
MTG-Jamendo	3768.9 hrs	55.6K	0.1	-
MusDB-HQ	29.1 hrs	10.2K	1.0	-
FMA	860.7 hrs	104.2K	0.4	0.1

B. Generated dialogue datasets

B.1. Overview

In this section, we introduce methods to generate our `AF-Dialogue-AudioSetSL` and `AF-Dialogue-MusicCaps` datasets with GPT-4 (Achiam et al., 2023). `AF-Dialogue-AudioSetSL` is generated based on the annotated events and timestamps of strongly labeled AudioSet-SL (Gemmeke et al., 2017; Hershey et al., 2021). There are 76k dialogues in the train split and 1.5k dialogues in test split. `AF-Dialogue-MusicCaps` is generated based on tags and descriptions of MusicCaps (Agostinelli et al., 2023). There are 3.3k dialogues in the train split and 1.4k dialogues in the test split. Each training sample has one to four rounds (QA pairs), and each testing sample has three or four rounds.

B.2. Prompts

Below we show the prompts used to generate audio-based multi-turn dialogues from GPT-4 for the strongly labeled AudioSet-SL dataset (Gemmeke et al., 2017; Hershey et al., 2021) and the MusicCaps dataset (Agostinelli et al., 2023) respectively.

sys prompt: Based on the sound events, create a dialogue between you (the assistant) and a person (the user) about the events in the audio. Each dialogue should consist of:
1. A user examines the audio, and sends a reasonable and creative message to the assistant. 2. Once the audio is provided, the assistant thoroughly perceives and comprehends them, responding with helpful answers that provide comprehensive reasoning. Do not include timestamps in the answer provided by the assistant. 3. Considering the past dialogue i.e. the question and the answer in the previous timestep, the user chooses another question or further inquiry. 4. The assistant provides a helpful answer providing comprehensive reasoning regarding the audio content.
The audios are 10 seconds long and the sound events are segmented based on where they occur in the audio. All the questions asked by the user should be diverse and complex.

Example 1: Sound Events: Sound of Ambulance (siren) (High-pitched and wailing): [0.0s-1.0s]; Sound of Traffic noise, roadway noise (Droning, loud and intrusive): [0.0s-10.0s]; Sound of Accelerating, revving, vroom (Highpitched, short and intense): [2.0s-10.0s]; Sound of Generic impact sounds (Loud and sharp): [6.7s-6.8s].
Dialogue:
User: What does the audio sound like?
Assistant: An ambulance siren echoes while traffic noise fades, and an engine revs.
User: Can you describe in more detail about the engine revs?
Assistant: The engine revving is highpitched, short and intense.
User: From the highpitched sound, can you tell me what vehicle it was?
Assistant: It is hard to tell this from the information provided.
User: Are there any other events in the audio that might be relevant?
Assistant: There is a sound of impact which is very loud and sharp.

Example 2: Sound Events: Sound of Outside, rural or natural: [0.000s-10.000s]; Sound of Liquid: [0.474s-0.680s], [4.249s-5.292s]; Sound of Gurgling, bubbling: [0.474s-0.680s], [4.249s-5.292s]; Sound of Splash, splatter: [1.620s-3.116s]; Sound of Laughter: [2.745s-4.029s], [5.299s-10.000s]; Sound of Chuckle, chortle: [2.752s-4.029s]; Sound of Speech: [4.551s-4.935s]; Sound of Male speech, man speaking: [4.557s-4.935s]; Sound of Belly laugh: [5.292s-10.000s]
Dialogue:
User: What are the main events in the audio?
Assistant: Liquid sounds, laughter, speech, and a man speaking can be heard in a rural setting.
User: Is is he speaking to other people?
Assistant: As there is sound of laughter present from other people, the man might be speaking to others.
User: Can you explain in more detail about the first event?
Assistant: The liquid sounds are more of splashing, thrashing and splattering sounds.
User: Are there other people laughing with him?
Assistant: Yes, there are other laughter voices along with that of the man.

Based on the above examples, design a four-turn dialogue between the user and assistant for the example given below. The question asked by the user should use pronouns such as ‘he’, ‘she’, ‘it’, ‘they’ and ‘them’ while referring to entities that have already occurred in the past questions or answers. Format each QA pair in a single line as a JSON dictionary (key “user” for question, and “assistant” for answer, wrapped with { and }). Do not include any other explanation.
Example 3:

sys prompt: Based on the music, create a dialogue between you (the assistant) and a person (the user) about the events in the music audio. Each dialogue should consist of:
1. A user examines the music, and sends a reasonable and creative message to the assistant. 2. Once the music audio is provided, the assistant thoroughly perceives and comprehends them, responding with helpful answers that provide comprehensive reasoning. Do not include timestamps in the answer provided by the assistant. 3. Considering the past dialogue i.e. the question and the answer in the previous timestep, the user chooses another question or further inquiry. 4. The assistant provides a helpful answer providing comprehensive reasoning regarding the music content. The more complex and diverse the question, the better. You can use the list of questions below while asking questions.
List of questions: Is it a live recording or a studio recording?
Is the guitar acoustic or electric?
Is it playing chords, a riff or a solo?
What music style or genre is it?
What key is it in?
How about the time signature?
What's the tempo? Slow, Medium or Fast?
Does it have vocals?
Is it low pitched or high pitched?
Is the person speaking fast or slow?
What is the language and accent?
What emotion is conveyed by the voice?
What are the vocal characteristics of that emotion?
How old is the person likely to be given their vocal characteristics?

Example 1:
Music description: House music with a thought provoking message in the beginning that eventually builds up and can be used to create the sense of a social revolution.
Dialogue:
User: What type of music is this?
Assistant: It is house music with synth and drums and a man speaking at the beginning.
User: What is his tone while speaking?
Example 2: He has an intense and energetic tone in his voice.
User: Does he speak till the end?
Assistant: No, he stops speaking halfway and then it is followed by house music.
User: What is the source of this music?
Assistant: This is electronic music.
Example 2:
Music description: Energetic bluesy song with a harmonica and horn section in musical dialogue.
Dialogue:
User: What instruments are playing in this music?
Assistant: This is a jazzy tune with harmonica, trumpet and drums playing.
User: Which key is the music playing in?
Assistant: The key is A Major.
User: How will you explain the tempo of this piece?
Assistant: The music is bouncy and positive with high beats per minute in the beginning and then changes signature halfway through.
User: What is the source of this music?
Assistant: This is fusion jazz as it has elements of bass and synthesizer.
Based on above examples, design a four-turn dialogue between the user and assistant for the example given below. The question asked by the user should use pronouns such as 'he', 'she', 'it', 'they' and 'them' while referring to entities that have already occurred in the past questions or answers. Format each QA pair in a single line as a JSON dictionary (key "user" for question, and "assistant" for answer, wrapped with { and }). Do not include any other explanation.
Example 3:

B.3. Dialogue filtering

The dialogues generated by GPT-4 as discussed in B.2 do not always follow the prompts, resulting in answers that have phrases such as "does not specify", "cannot be determined", "without additional context" and so on. Hence, following Gardner et al. (2023), we filter such QA pairs to improve the data quality and ensure desirable outputs from the model.

Apart from the manual filtering step, we also filter samples based on the similarity of the answer generated by GPT-4 and the audio samples. Specifically, we compute the cosine similarity between the LAION-CLAP text-embeddings and audio-embeddings (Wu et al., 2023) for a given QA pair in each dialogue. The distributions of similarities are shown in Figure 8. We remove samples if the similarity is below a threshold of 0.3.

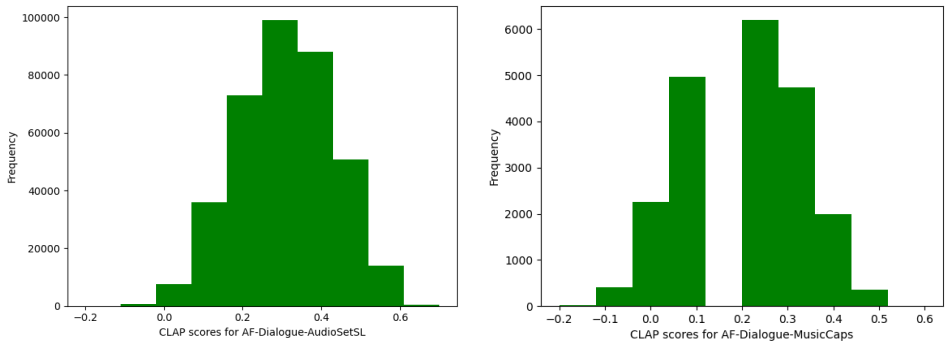


Figure 8. Distributions of LAION-CLAP similarities before filtering the generated `AF-Dialogue-AudioSetSL` (left) and `AF-Dialogue-MusicCaps` (right) datasets.

B.4. Examples of generated dialogues

We demonstrate an example of generated dialogue for each of the generated dataset.

```
[Audio annotations:  Sound of Mechanisms (Produces sound through mechanical means.):
[0.119s-10.000s]; Sound of Generic impact sounds (Short, sharp, and loud bursts.):
[0.217s-0.387s], [2.394s-2.727s], [2.889s-3.255s], [3.799s-4.132s], [4.473s-4.766s],
[4.928s-5.188s], [8.129s-10.000s]; Sound of Glass shatter (High-pitched, sharp, sudden,
loud crash.):  [1.622s-4.652s]; Sound of Walk, footsteps (Soft, rhythmic, echoing thuds.):
[5.724s-6.106s], [6.455s-6.650s], [7.024s-7.479s], [7.625s-7.901s]]
```

User: Can you describe the main events in the audio?
Assistant: The audio mainly consists of mechanical sounds, sharp impact sounds, the shattering of glass, and the sound of footsteps.
User: What can you tell me about the impact sounds?
Assistant: The impact sounds are short, sharp, and loud bursts. They occur at several points throughout the audio, suggesting some kind of repetitive action or event.
User: What about the glass shattering, does it happen more than once?
Assistant: The glass shattering sound is a high-pitched, sharp, sudden, loud crash. It occurs once in the audio, lasting for about three seconds.

```
[Audio annotations:  This is the recording of a cowbell.  The cowbell is being played with
the side and the tip of the drumstick repeatedly.  This recording could be sampled for use
in beat-making especially for genres that have a heavy usage of cowbells such as latin dance
music.]
```

User: What instrument is being played in this audio?
Assistant: The instrument being played in this audio is a cowbell.
User: How is it being played?
Assistant: The cowbell is being played with the side and the tip of the drumstick repeatedly.
User: What is the tone of the cowbell?
Assistant: The tone of the cowbell is bright, resonant, and penetrating.

C. Additional Experiments, Examples and Discussion of Audio Flamingo

We refer readers to visit our demo website: `https://audioflamingo.github.io/` for comprehensive demos.

C.1. Additional Few-shot Experiments on Unseen Labels

Audio Flamingo can be adapted to classification with completely unseen labels with the help of ICL-based RAG. In Table 9 we demonstrate Audio Flamingo can be adapted to new classification tasks where none of the labels have been seen during training. The labels are bird species and gun names, respectively. The huge improvement from zero-shot to few-shot indicates Audio Flamingo's strong ability to follow retrieved in-context samples.

Table 9. Accuracy of Audio Flamingo on new classification tasks with unseen labels.

Dataset	Zero-shot	Few-shot
`BG-Gun-Sound` (Park et al., 2022)	1.6%	53.5%
`Birds-Dataset-Till-2022` (Mohanty)	0.0%	42.2%

C.2. Compute

In this section we compare the computation resources required by our model and baseline models. Pengi (Deshmukh et al., 2023) is based on GPT2-base and has a total size 308M. It's trained on 20 V100 GPUs. LTU (Gong et al., 2023c) is based on LLaMA-7B and has a total size 6.74B. It's trained on 4 RTX A6000 GPUs for about 3 days. Qwen-Audio (Chu et al., 2023) is based on Qwen-7B and has a total size of 8.3B. The training cost is unknown. Salmonn (Tang et al., 2023a) is based on Vicuna-13B. The training cost is unknown. Our proposed Audio Flamingo has a total size of 2.2B. We use 8 A100 GPUs. It takes about 5-6 days for pretraining, about 3 days for SFT, and about 2hrs for dialogue fine-tuning.

C.3. Audio Flamingo understands sound in depth

Audio Flamingo understands occurrences in audio.
`[Subjective description of the audio: three people talking in order, with a lot of noise in the recording.]`
Task: *Audio Question Answering*
Instruction: *How many different voices are there?*
Ground Truth: *Three.*
Baseline: *One.*
Audio Flamingo: *Three.*

Audio Flamingo understands the order of sounds.
`[Subjective description of the audio: a man speaks on a talk show, then the audience laugh and clap their hands.]`
Task: *Audio Captioning*
Instruction: *Describe this sound.*
Ground Truth: *A man is speaking followed by applause.*
Baseline: *A crowd of people are speaking and applauding.*
Audio Flamingo: *A man is speaking and then the audience claps.*

Audio Flamingo understands quality of audio.
`[Subjective description of the audio: a beautiful and sad piece of instrumental music starting with keyboard playing chords slowly followed by soft violin.]`
Task: *Audio Question Answering*
Instruction: *What is the atmospheric quality of the audio?*
Ground Truth: *The atmospheric quality of the audio is its ability to create a sense of space and depth, with a focus on creating a sense of mystery and wonder.*
Baseline: *The audio has a rich and dynamic quality, which suggests a lively and engaging atmosphere, possibly associated with music or a musical performance, but it is not specified in the given audio events.*
Audio Flamingo: *The audio has a melancholic and introspective atmosphere.*

Audio Flamingo perceives loudness and its change over time.
`[Subjective description of the audio: a man speaks in a small volume, then there is very loud burst, and then the man laughs for very short time in a surprised way.]`
Task: *Audio Captioning*
Instruction: *Describe the sound in a sentence.*
Ground Truth: *A man briefly talks followed by a loud burst then laughter.*
Baseline: *A person is laughing and a gun is fired.*
Audio Flamingo: *A man is speaking and then a loud bang occurs.*

Audio Flamingo understands distances of sounds.
`[Subjective description of the audio: a man speaks in the rain, then another man laughs very slightly, followed by the sound of thunderstorm.]`
Task: *Audio Captioning*
Instruction: *Describe this sound.*
Ground Truth: *A man talking followed by another man laughing as rain falls and thunder roars in the distance.*
Baseline: *A man is talking and rain is falling.*
Audio Flamingo: *A man is talking in the rain with thunder in the background.*

C.4. Audio Flamingo looks at retrieved samples

Audio Flamingo borrows key words from retrieved samples.
Task: *Audio Captioning*
Instruction: *Describe this sound.*
Ground Truth: *Paper is crinkling and crumpling, and thumping occurs.*
Baseline: *Movement of a person in a kitchen with pots and pans on a counter.*
Retrieved samples:
1) Some material is crumpled.
2) Plastic is being crumpled.
3) Crinkling of a paper with a small step.
Audio Flamingo: *A person crumples up a sheet of paper and crinkles it.*

Task: *Audio Captioning*
Instruction: *Describe this sound.*
Ground Truth: *a woman speaks, and faucet water runs.*
Baseline: *A woman is talking to a child.*
Retrieved samples:
1) A woman speaking and water running.
2) A women voice followed by flowing water from a faucet.
3) Water sprays from a faucet, a person speaks.
Audio Flamingo: *A woman speaks while water runs from a faucet.*

Audio Flamingo can ignore noisy retrieved samples and make correct prediction.
Task: *Music Genre Classification*
Instruction: *What is the genre of this music?*
Ground Truth: *pop.*
Retrieved samples:
1) reggae.
2) jazz.
3) disco.
Audio Flamingo: *pop.*

Task: *Music Instrument Classification*
Instruction: *What is the instrument of this music?*
Ground Truth: *flute.*
Retrieved samples:
1) clarinet.
2) clarinet.
3) clarinet.
Audio Flamingo: *flute.*

C.5. Audio Flamingo has strong dialogue ability

```
[Audio annotations:  Sound of Chirp, tweet (High-pitched, short, repetitive bird sounds.):
[0.000s-1.157s], [1.543s-2.055s], [3.220s-4.409s], [4.748s-5.709s], [5.921s-6.780s],
[7.094s-7.976s], [8.331s-8.819s], [9.142s-10.000s]; Sound of Animal (Loud, diverse, and
often rhythmic.):  [0.000s-1.157s], [1.528s-2.031s], [3.220s-4.386s], [4.764s-5.701s],
[5.929s-6.740s], [7.071s-7.976s], [8.331s-8.795s], [9.134s-10.000s]; Sound of Bird vocalization,
bird call, bird song (Melodic, rhythmic, varied pitch and tone.):  [0.000s-1.165s],
[1.528s-2.039s], [3.213s-4.386s], [4.772s-5.693s], [5.921s-6.701s], [7.047s-7.953s],
[8.331s-8.780s], [9.150s-10.000s]; Sound of Bird (High-pitched, melodic, rhythmic chirping
sounds.):  [0.000s-1.173s], [1.535s-1.992s], [3.205s-4.362s], [4.780s-5.669s], [5.906s-6.669s],
[7.016s-7.929s], [8.323s-8.772s], [9.134s-10.000s]; Sound of Speech (S̈ound waves produced
by vocal cords.̈):  [2.024s-3.008s], [4.630s-5.283s], [6.835s-8.134s], [8.701s-10.000s];
Sound of Female speech, woman speaking:  [2.047s-3.039s], [4.614s-5.315s], [6.843s-8.150s],
[8.709s-10.000s]]
```

User: *Can you briefly explain what you hear in the audio?*
Ground Truth: *The audio primarily consists of bird sounds, including chirping and bird vocalizations. There are also sounds of other animals. Additionally, there is a woman speaking at certain intervals.*
Baseline: *Birds are chirping.*
Audio Flamingo: *The audio features a variety of sounds including bird vocalizations, mechanical sounds, impact sounds, a man speaking, and a ticking sound.*

User: *Are the bird sounds continuous?*
Ground Truth: *The bird sounds are not continuous. They occur at different intervals throughout the audio, with some overlap with the woman's speech.*
Baseline: *Yes, the bird sounds are continuous.*
Audio Flamingo: *No, the bird sounds are not continuous. They occur at different intervals throughout the audio.*

User: *Do the bird sounds and the person's speech overlap?*
Ground Truth: *Yes, there are instances where the bird sounds and the woman's speech overlap.*
Baseline: *Yes, the bird sounds and the woman's speech overlap.*
Audio Flamingo: *Yes, the bird sounds and the person's speech do overlap at certain points in the audio.*

`[Audio annotations: Sound of Music (Sound produced by vibrating instruments.): [0.000s-10.000s]; Sound of Rapping (Rhythmic speech over beats.): [0.000s-10.000s]; Sound of Male singing (Deep, resonant, and powerful tones.): [0.000s-10.000s]]`

User: *Is the male voice rapping or singing? Ground Truth: The male voice is doing both, rapping and singing in the audio.*
Baseline: *The male voice is rapping.*
Audio Flamingo: *The male voice is rapping.*

User: *What kind of music is it?*
Ground Truth: *The type of music isn't specified, but it's accompanied by rapping and male singing.*
Baseline: *The kind of music is hip-hop.*
Audio Flamingo: *The music is a rap song.*

User: *Is the male voice the only one present?*
Ground Truth: *Yes, the male voice is the only one specified in the audio, both rapping and singing.*
Baseline: *Yes, the male voice is the only one present.*
Audio Flamingo: *Yes, the male voice is the only one present in the audio.*

`[Audio annotations: Sound of Music (Sound produced by vibrating instruments.): [0.000s-10.000s]; Sound of Male singing (Deep, resonant, and powerful tones.): [1.341s-3.734s], [9.289s-10.000s], ]`

User: *What genre does this music belong to?*
Ground Truth: *This music belongs to the R&B/Disco genre.*
Baseline: *It is not specified in the given audio clip.*
Audio Flamingo: *This music belongs to the R&B genre.*

User: *Can you tell me more about the vocals in this track?*
Ground Truth: *Sure, the vocals in this track are from a male singer who is singing in a higher register. His voice has been modified with effects, making it sound like a voice sample.*
Baseline: *The vocals in this track are described as rich and full-bodied, but no further details are given about the specific singer or the lyrics being sung in the audio clip.*
Audio Flamingo: *The vocals in this track are by a male singer. His voice is deep, resonant, and powerful, adding to the overall emotional impact of the song.*

Exhibit D

Jamendo Licensing

Jamendo S.A.
51 Rue de Strasbourg
2561 - Luxembourg
Luxembourg
VAT: LU22234927

NVIDIA Corporation
2788 San Tomas Expressway
Santa Clara CA 95051
United States

Invoice S0002/2025/00380 - PF

Invoice Date
23/09/2025

Due Date
23/09/2025

DESCRIPTION	QUANTITY	UNIT PRICE	TAXES	AMOUNT
License - International	55.600,00 Units	289,0000	0%	16.068.400,00 €

Conditions de règlement :Immediate Payment

Payment Communication: **RF84 9191 87**
on this account: **LU42 0141 4365 1810 0000 - ING**

This invoice is subject to the general terms and conditions of sale available on the website of Jamendo, which have been accepted by the client by downloading the tracks.

Untaxed Amount	16.068.400,00 €
TVA 0%	0,00 €
Total	**16.068.400,00 €**

Exhibit E

December 16, 2025

Via FedEx

Jamendo, S.A.
51 Rue de Strasburg
2561 – Luxembourg
Luxembourg

Re: Invoice S0002/2025/00379-PF

Dear Sir or Madam,

I am writing to you on behalf of NVIDIA Technologies Belgium BV ("NVIDIA Belgium") regarding the above-referenced invoice from Jamendo SA dated September 23, 2025, and the corresponding notice of default dated October 14, 2025.

We are aware that legal proceedings are currently pending related to this invoice. The following remarks are provided solely for the sake of completeness.

The invoice requests a 16,068,400.00 € payment for a "license" of "55.600,00 Units" at a "289,0000" unit price. Our employees in Belgium initially assumed that this invoice was sent by mistake or that it was a phishing attempt. Not only is the invoice marked as "pro forma" (PF), it also does not reference a corresponding purchase order from NVIDIA Belgium or a related business transaction between NVIDIA Belgium and Jamendo. In fact, the invoice doesn't even describe the "Units" that are allegedly being licensed, except for a mention in passing of "the client . . . downloading the tracks." The notice of default likewise fails to reference a corresponding purchase order or related business transaction other than "downloading of the tracks from [Jamendo's] website."

We have confirmed that there is no purchase order or other business transaction between NVIDIA Belgium and Jamendo. As a result, and in accordance with our accounting policies and procedures, NVIDIA Belgium rejects and will not pay the invoice.

The assertion of claims against your company due to unauthorized and/or fraudulent invoicing is expressly reserved.

Thank you for your attention to this matter.

Sincerely,

Rebecca Peters

DocuSigned by:
Rebecca Peters
7AF680BB99BB4A1...

Director
NVIDIA Technologies Belgium BV

17 December 2025

Exhibit F

Certificate of Registration

This Certificate issued under the seal of the Copyright Office in accordance with title 17, *United States Code*, attests that registration has been made for the work identified below. The information on this certificate has been made a part of the Copyright Office records.

Registration Number
TX 9-606-566
Effective Date of Registration:
June 16, 2026
Registration Decision Date:
June 17, 2026

Title

Title of Work: MTG-Jamendo Dataset

Completion/Publication

Year of Completion: 2019
Date of 1st Publication: June 15, 2019
Nation of 1st Publication: Spain

Author

- **Author:** Jamendo S.A.

Author Created: compilation of data
Work made for hire: Yes
Citizen of: Luxembourg
Domiciled in: Luxembourg

Copyright Claimant

Copyright Claimant: Jamendo S.A.
51, Rue de Strasbourg, Luxembourg, 2561, Luxembourg

Rights and Permissions

Organization Name: Jamendo S.A.
Name: Olivier Van Gulck
Email: olivier.vangulck@winamp.com
Telephone: +32475213755
Address: 51, Rue de Strasbourg
Luxembourg 2561 Luxembourg

Certification

Name: Breane Stryker
Date: May 06, 2026

Correspondence: Yes