

AURELIUS: RELATION AWARE TEXT-TO-AUDIO GENERATION AT SCALE

005 **Anonymous authors**

006 Paper under double-blind review

ABSTRACT

011 We present *Aurelius*, a new framework that enables relation aware text-to-
 012 audio (TTA) generation research at scale. Given the lack of essential audio event
 013 and relation corpora, *Aurelius* contributes a large-scale audio event corpus *AudioEventSet* and another large-scale relation corpus *AudioRelSet*. Comprising
 014 110 event categories, *AudioEventSet* maximally covers all commonly heard audio
 015 events and each event is unique, realistic and of high-quality. *AudioRelSet* con-
 016 sists of 100 relations, comprehensively covering the relations that present in the
 017 physical world or can be neatly described by text. As the two corpora provide
 018 audio event and relation independently, they can be combined to create massive
 019 $\langle \text{text}, \text{audio} \rangle$ pairs with our pair generation strategy to support relation aware
 020 TTA investigation at scale. We comprehensively benchmark all existing TTA
 021 models from both general and relation aware evaluation perspective. We further
 022 provide in-depth investigation on scaling up existing TTA models' relation aware
 023 generation by either training from scratch or leveraging cross-domain general TTA
 024 knowledge. The introduced corpora and the findings through investigation in this
 025 work potentially facilitate future research on relation aware TTA generation.

1 INTRODUCTION

029 Text-to-audio (hereinafter TTA) generation task aims at generating acoustically high-fidelity audio
 030 with the content inferred by the input text. Owing to the success of generative modeling (e.g.,
 031 diffusion based (Ho et al., 2020; Xue et al., 2024), score based (Vahdat et al., 2021) and flow
 032 matching based (Lipman et al., 2023; Guan et al., 2024) methods) and the availability of large
 033 $\langle \text{text}, \text{audio} \rangle$ pair dataset (e.g., AudioCaps (Kim et al., 2019), AudioSet (Gemmeke et al.,
 034 2017)), we have witnessed significant advancement in general TTA task in recent years (Ghosal
 035 et al., 2023; Hung et al., 2024; Liu et al., 2024). Despite these achievements, the relation aware
 036 TTA generation still remains as a challenging task as it jointly requires audio event generation and
 037 relation modeling. Audio events and their relation are two fundamental elements humans rely on for
 038 holistic acoustic scene understanding or engaging communication (Zacks et al., 2007; Hirsh et al.,
 039 1967; Lake et al., 2015). We humans can interpret the relation and audio events within the textual
 040 description with ease to decide how the target audio looks like. Enabling TTA models with similar
 041 relational reasoning and event interpretation capability is therefore essential for bridging the gap
 042 between relation aware TTA model quality and human-level crossmodal reasoning.

043 The recent preliminary investigation by RiTTA (He et al., 2025) already shows the incapability of
 044 existing TTA models in relation aware generation, but the investigation runs on top of small relation
 045 and audio event corpora. The data corpora small scale issue naturally hinders further investigation.
 046 To enable relation aware TTA at scale, we introduce *Aurelius*, a novel framework that contributes
 047 to relation aware TTA from both dataset benchmark and technical methodology aspects. From
 048 the dataset benchmark aspect, we meticulously curate two large-scale corpora: *AudioEventSet* and
 049 *AudioRelSet*. *AudioEventSet* is an audio event corpus that comprises 110 across fine-grained event
 050 classes across 7 main acoustic categories we commonly hear in our daily lives. In contrast to existing
 051 audio event datasets (Gemmeke et al., 2017; Kim et al., 2019; Fonseca et al., 2022) that are either
 052 noisy, polyphonic or label-missing, *AudioEventSet* provides a coarse-to-fine tree structured audio
 053 event corpus that is both internally distinctive and externally comprehensive. Each individual audio
 event in *AudioEventSet* is high-quality, realistic and intra-class diverse. *AudioRelSet* is the large-scale
 relation corpus with up to 100 detailed relations completely covering the potential relations audio

events may present in the 3D physical world or text can describe succinctly. *AudioRelSet* is also tree structured and can be further scaled up to incorporate more relations. Each relation in *AudioRelSet* has an “arity” property that is further used to combine relation and audio events together to create $\langle \text{text}, \text{audio} \rangle$ pairs for relation aware TTA task. *AudioEventSet* and *AudioRelSet* are orders of magnitude larger than existing relevant dataset, enabling thorough and in-depth investigation for relation aware TTA task.

Based on the introduced audio event corpus *AudioEventSet* and relation corpus *AudioRelSet*, we further introduce a $\langle \text{text}, \text{audio} \rangle$ pair generation strategy that is capable of generating essential $\langle \text{text}, \text{audio} \rangle$ pairs highlighted by both audio event based and textual description diversity. As the audio event corpus is disentangled from relation corpus, our proposed strategy can generate nearly unlimited $\langle \text{text}, \text{audio} \rangle$ pairs tailored for various training requirements. In summary, as illustrated in Fig. 1, *Aurelius* advances relation aware TTA research by contributing large-scale corpora of audio events and relations, together with a dedicated framework for relation aware generation. The explicit disentanglement of audio events and relations, the hierarchical tree-structured design of each corpus, and the systematic $\langle \text{text}, \text{audio} \rangle$ creation strategy collectively provide a strong foundation for curating essential datasets in this domain. Building on this foundation, our proposed *AudioRelGen* framework tackles relation aware TTA by decoupling audio event modeling from relation modeling, offering an essential first step toward structured audio generation. We believe this work will not only establish a new benchmark for relation aware TTA but also inspire future research on modeling complex event–relation dynamics in sound.

2 RELATED WORK

Text-to-Audio Generation aims at generating the audio waveform that semantically aligns well with the input text. The fast development of generative modeling techniques (Ho et al., 2020; Vahdat et al., 2021; Lipman et al., 2023) in recent years has largely advanced the TTA generation in terms of high-fidelity and high-intelligibility (Liu et al., 2024; 2023; Kreuk et al., 2023; Yang et al., 2022; Ghosal et al., 2023; Liao et al., 2024), alongside other crossmodal generation tasks including but not limited to text-to-music (TTM, e.g., MusicGen (Copet et al., 2023) and MusicLM (Agostinelli et al., 2023)), image-to-audio (I2A, e.g., RegNet (Chen et al., 2020), Img2Wav (Sheffer & Adi, 2023) and SpecVQGAN (Iashin & Rahtu, 2021)) and text-to-image (T2I). Although the promising achievement in generating realistic and semantically text-aligned audio, existing TTA methods still perform poorly in relation aware TTA generation. Prior work like RiTTA (He et al., 2025) and CompA (Ghosh et al., 2024) have preliminarily explored relation aware TTA and shown the incapability of existing TTA methods through limited audio event and relation corpora, which inevitably hinders future investigation at scale. Moreover, publicly available audio event corpora (AudioSet) are directly collected from either online video data or audio sharing platform without proper quality check, resulting in the audio events label-missing, noisy and ambiguous. Our work circumvents these barriers by introducing a meticulously curated audio event corpus *AudioEventSet* that is of high-quality, distinctive and realistic, potentially covering all commonly heard audio events.

Relation Modeling has been widely discussed within modalities, including image (Liu et al., 2022; Zerroug et al., 2022), natural language processing (Wadhwa et al., 2023) and acoustics (Xie et al., 2025a; Ghosh et al., 2024; He et al., 2025). In the context of 2D image, the objects of interest can exhibit compositional and spatial relation (Liu et al., 2022; Zerroug et al., 2022). In the context of 3D physical world, audio event is the most fundamental acoustic signal and multiple audio events join together to represent the 3D physical world via more sophisticated relations than image-based relations, ranging from basic spatial, temporal, perceptual relation to their nested combination. Prior works (Xie et al., 2025a; Ghosh et al., 2024; He et al., 2025; Xie et al., 2025b) have discussed

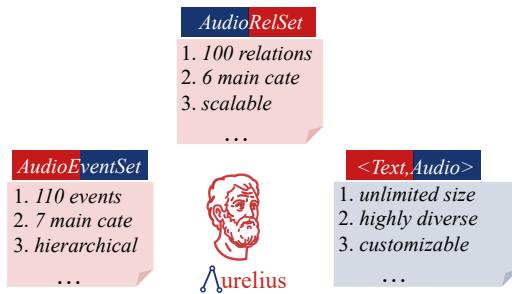


Figure 1: *Aurelius* contributes to relation aware TTA by introducing an audio event corpus *AudioEventSet*, a relation corpus *AudioRelSet* and $\langle \text{text}, \text{audio} \rangle$ pair generation strategy.

108 audio event relations in small-scale and with minimal complexity, making them hard to scale up to
 109 accommodate the potential relation complexity that present in either 3D physical environment or
 110 textual description. To fill in this gap, we curate *AudioRelSet*, a large-scale relation corpus that reflect
 111 the relation potentially present in the physical world and can be neatly describe by text.

112 **Text-to-Audio Generation Techniques.** Existing TTA methods can be technically divided into two
 113 main categories: while the early methods are diffusion based (Liu et al., 2024; 2023; Kreuk et al.,
 114 2023; Yang et al., 2022; Ghosal et al., 2023; Liao et al., 2024; Xue et al., 2024), the latest methods are
 115 flow-matching based (He et al., 2025; Hung et al., 2024; Guan et al., 2024). The flow-matching based
 116 methods are usually faster during both training and inference, and can give better performance than
 117 diffusion based methods. We completely benchmark all these methods on our introduced corpora,
 118 and further provide in-depth investigation to reveal potential ways scale up existing TTA methods'
 119 relation aware TTA capability.

120

121

3 AURELIUS BENCHMARK: AUDIOEVENTSET AND AUDIORELSET

122

123

3.1 AUDIO EVENT CORPUS: *AudioEventSet*

124

125

An audio event refers to an auditory signal occurring over a specific period of time, typically representing an independent, human-recognizable sound. To support the relation aware TTA research, the desired audio event corpus should be:
 1. diverse enough so as to maximally accommodate the wide variety of audio events potentially present in the 3D physical world; 2. clean and of high-fidelity so as to enable reliable in-depth technical investigation; 3. distinctive so that they can be easily distinguished without any ambiguity; 4. hierarchically organized w.r.t. their genre so as to enable investigation at different granularity. After thorough investigation on existing audio event related dataset, however, we find all existing datasets fall short in exhibiting the four properties. As is shown in Table 1, existing audio event dataset (e.g., AudioSet (Gemmeke et al., 2017), AudioCaps (Kim et al., 2019), AudioTime (Xie et al., 2025a) and FSD50K (Fonseca et al., 2022)) are either noisy, label-missing, polyphonic (multiple events temporally overlap) or semantically ambiguous (where multiple event classes correspond to the same audio). To address this dilemma, we introduce *AudioEventSet*, a meticulously curated audio event corpus that is intrinsically clean, diverse, distinctive and hierarchically organized.

126

127

AudioEventSet ontology is tree-structured and the tree depth is three. From the root node to the leaf node, each audio event is organized in coarse-to-fine granularity. As is shown in Fig. 2 and Table I in Appendix, we base on RiTTA (He et al., 2025) to categorize *AudioEventSet* into seven main categories: five singular-source categories *Animal*, *Human*, *Machinery*, *Music* and *Nature*, two interaction-based categories *Human-Object* and *Object-Object* interactions. The seven categories maximally cover the commonly heard audio events in the 3D physical world. Each main category associates with multiple subcategories, each of which is further associated with multiple fine-grained event classes. For example, the *Human* main category contains *human voice*, *human speech*, *hands action*, *group action* and *locomotion* subcategories, comprehensively categorizing the human centered audio event from various aspects.

128

129

During *AudioEventSet* ontology construction, we guarantee each curated audio event is distinctive, unique, and human-distinguishable. Audio event emitting ambiguous or nondistinctive audio is discarded. For example, *engine idling* in AudioSet (Gemmeke et al., 2017) audio differs significantly by various engines, and it easily confuses with another audio event such as *working fan* and *hairdryer*. We thus exclude all of them from the corpus. Moreover, we account for both audio event source origin, event category and the audio generation physical mechanism for *AudioEventSet* ontology construction. For example, in the Object-Object main category, we exhaustively consider the impact, friction, dropping and explosion audio generation mechanism. In summary, we have curated 110

Table 1: Audio Event Dataset Comparison.

Dataset	Characteristic
AudioSet (2017)	<i>polyphonic, ambiguous, noisy, label-missing</i>
FSD50K (2022)	
AudioCaps (2019)	
AudioTime (2025a)	
<i>AudioEventSet</i>	<i>distinctive, high-quality, clean, hierarchical coarse-to-fine intra-class diversity inter-class discriminative</i>

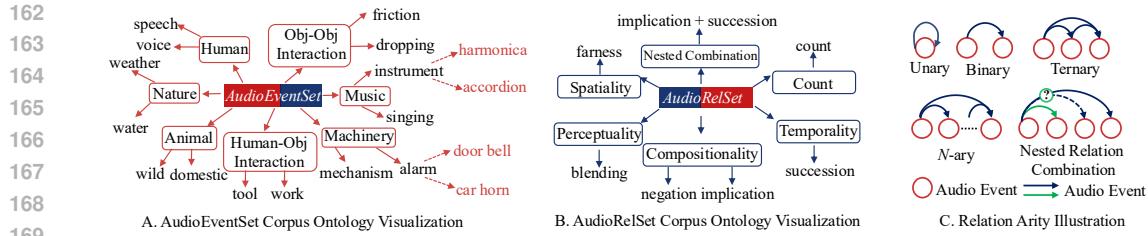


Figure 2: *AudioEventSet* and *AudioRelSet* corpora illustration: we visualize the *AudioEventSet* ontology in sub-figure A. It is tree-structured with depth 3 and contains 7 main categories and 110 event categories (leaf node) in total. We just show part of the leaf nodes (with red dotted arrow) for the sake of clear visualization. The detailed event ontology is given in Table I in Appendix. The *AudioRelSet* ontology in sub-figure B, it is tree-structured with depth 2. It contains 6 main categories and 100 categories in total. The detailed relation ontology is given in Table II in Appendix. In sub-figure C, we conceptually illustrate the relation “arity”, which is used to connects relation and audio event to generate audios.

audio events, which is four times larger than audio event corpus proposed in RiTTA (He et al., 2025), each leaf audio event is associated with around 75 realistic audio snippets ranging from 1 s to 5 s

For each leaf node audio event, we collect exemplar audios from either copyright-free freesound.org platform or FSD50K (Fonseca et al., 2022). As most audios from freesound.org and FSD50K¹ real audios shared by volunteers across the globe, the collected audios for each audio event are diverse and realistic enough to reflect the audio event we can hear in the physical world. Manually verification is adopted to ensure the collected exemplar audios content correctness, label consistency. We argue that the curated *AudioEventSet* can be potentially applied to other tasks other than TTA, we anticipate much wider usage of the curated dataset.

3.2 RELATION CORPUS: *AudioRelSet*

Prior works (Xie et al., 2025a; He et al., 2025; Ghosh et al., 2024) have explored audio events relation from various perspectives, but only on a small scale. For example, AudioTime (Xie et al., 2025a) and CompA (Ghosh et al., 2024) have discussed temporal relations. RiTTA (He et al., 2025) has additionally introduced spatial, compositional and count relations, resulting in a total of 11 relations. In this section, we introduce *AudioRelSet*, a meticulously curated large-scale relation corpus with up to 100 distinct relations. To ensure *AudioRelSet* to exhibit both real scenario practicability, text-manageable complexity and relation scalability, we follow 3 guidelines to curate *AudioRelSet*: 1. maximally cover the potential relations audio events can present in the 3D physical world; 2. enough relation complexity but can still be efficiently and neatly described by text; 3. the relation corpus can be scaled up to accommodate more sophisticated relations. To this end, we construct 6 main fundamental relations, in which 4 main relations describe the relations present in the 3D physical world, one main relation focuses on TTA model’s logical reasoning capability and last one relation derives from the nested combination of the five main relations.

AudioRelSet ontology is tree-structured and the tree depth is 2, the root node connects 6 main relations, each of which further associates with multiple sub-relations. Let $\mathcal{E} = \{E_1, E_2, \dots, E_m\}$ denote the audio events in *AudioEventSet* introduced in Sec. 3.1, $\mathcal{R} = \{R_1, R_2, \dots, R_n\}$ denote the relations to be constructed. *AudioRelSet* is represented as follows,

1. **Temporality** describes the sequence or overlap of audio events in time domain, it contains 4 sub-relations: *Precedence*: $E_1 < E_2$ (event E_1 occurs before E_2); *Succession*: $E_1 > E_2$ (event E_1 occurs before E_2); *Simultaneity*: $E_1 \parallel E_2$ (E_1 and E_2 occur concurrently); *Repetitiveness*: $\sim E_1$ (event E_1 occur repetitively in the time domain).
2. **Spatiality** defines the relative spatial positions or motion status between or within audio events, it contains 5 sub-relations: *Proximity*: $d(E_1, E_2) \leq \tau$ (E_1 E_2 are within distance τ); *Closeness*:

¹FSD50K (Fonseca et al., 2022) data is also sourced from freesound.org

216 $d(E_1) < d(E_2)$ (E_1 is closer than E_2); *Farness*: $d(E_1) > d(E_2)$ (E_1 is further than E_2); *Approaching*: $\frac{d}{dt}d_{E_1}(t) < 0$ (E_1 is moving close); *Departuring*: $\frac{d}{dt}d_{E_1}(t) > 0$ (E_1 is moving away).

219 3. **Count** focuses on the number of audio events take place within a period of time: *Count*: $|\mathcal{E}| = N, N \in \mathbb{Z}^+$. (cardinality \mathcal{E} is the number).

221 4. **Perceptuality** introduces 6 acoustic effects to an audio event,

223 • *Balancing*: $\mathcal{R}_{\text{balance}}(E_1, E_2, \sigma)$ (level balance between E_1 and E_2 by balancing factor σ , so that
224 one event dominates and the other serves as the background audio).

225 • *Blending*: $\mathcal{R}_{\text{blend}}(E_1, E_2, \theta)$ (mix E_1 and E_2 together by factor θ so as to be indistinguishable).

226 • *Reverberation*: $\mathcal{R}_{\text{reverb}}(E_1)$ applies reverberation effect to E_1 , as if it is heard in the canyon.

227 • *Time-stretching*: $\mathcal{R}_{\text{stretch}}(E_1, \alpha)$, where α is the time-stretching factor and E_1 listens slowly.

228 • *Amplification*: $\mathcal{R}_{\text{amp}}(E_1, \beta)$, where β is the amplification factor and E_1 listens to be louder.

229 • *Attenuation*: $\mathcal{R}_{\text{att}}(E_1, \gamma)$, where γ is the attenuation factor and E_1 listens to be quieter.

231 5. **Compositionality** indicates the logical operation within audio events TTA models need to reason
232 before deciding what audio events to generate. It contains 5 sub-relations.

234 • *Conjunction*: $E_1 \wedge E_2$ (both events occur).

235 • *Disjunction*: $E_1 \vee E_2$ (at least one event occurs, or both occur).

236 • *Negation*: $\neg E_1$ (the absence of the event E_1 in the generated audio).

237 • *Exclusive Or*: $(E_1 \vee E_2) \wedge \neg(E_1 \wedge E_2)$ (either E_1 or E_2 occur, but not both).

238 • *Implication*: $E_1 \Rightarrow E_2, \neg E_1 \Rightarrow E_3$ (if E_1 occur, then E_2 occur, else E_3 occur).

241 6. **Nested Combination** is a hierarchical structuring of multiple basic relations (e.g., the aforementioned
242 *Temporality*, *Spatiality*), such that the output of one relation serves as the input or context
243 for another, forming a directed acyclic relation structure. Nested combination allows for capturing
244 complex relation interactions among audio events. For example, by nesting *Implication*, *Approaching*
245 and *Conjunction*, we can generate a more complex text prompt showing below,

246 Nest Combination Example: *Implication*, *Approaching* and *Conjunction*

248 If generated both {A} event and {B} event, → *Conjunction*
249 then continue to generate {C} audio event,
250 else just generate {D} audio event that is gradually approaching close. → *Approaching*

252 Mathematically, the relation $R_{\text{nested}}(E)$ resulting from nested combination can be represented as,
253

$$R_{\text{nested}}(E) = R_n(R_{n-1}(\dots R_2(R_1(E)) \dots)) \quad (1)$$

254 where $E = \{e_1, e_2, \dots, e_m\}$ represents a finite
255 set of audio events. We combine relations arising
256 from the introduced 5 basic relations to construct
257 nested combination relations and have created
258 79 nested combination relations.

259 It is worth noting that the nested combination is
260 scalable and we can theoretically construct more
261 complex nested relations (even infinite relations)
262 by simply involving more basic relations into the
263 nested combination process. In this work, we
264 constrain the nested combination up to involving
265 5 audio events at most (*Quinary*), it remains as
266 future research topic to explore more complex
267 nested combination, and the key challenge re-
268 mains on how to construct the corresponding concise and precise textual description for the given

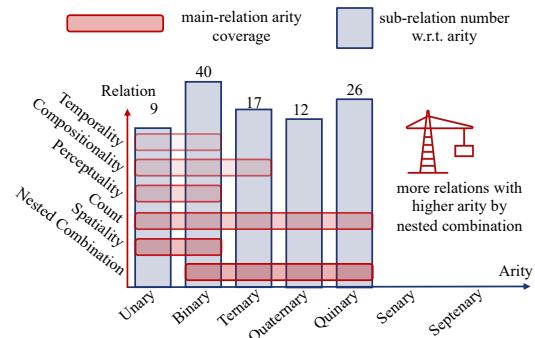


Figure 3: Arity coverage in *AudioRelSet*.

270 highly complex nested relation. Moreover, during the nested combination process, we explicitly run
 271 internal nested relations logic correctness and feasibility check before accepting the nested relations
 272 as a new relation, any nested relation violating the correctness and feasibility rule is abandoned. For
 273 example, the combination of *Count* and *Conjugation* internally equals to *Count*.

274 **Relation Arity.** each relation in *AudioRelSet* is associated with an “arity” property, which indicates
 275 the audio event number it requires to represent the relation. The visual illustration of arity is shown
 276 in Fig. 2 C. The arity coverage across *AudioRelSet* main relation categories is given in Fig. 3, from
 277 which we can see that the arity ranges from 1 to 5 (unary to quinary) and most main relation cuts
 278 across multiple arities. Moreover, the construction of more complex relations introduces higher arity.
 279 We use “arity” to create $\langle \text{Text}, \text{audio} \rangle$ pairs (see Sec. 3.3) and experiment evaluation (see Sec. 4).
 280

281 **3.3 TEXT-AUDIO PAIR CREATION: $\langle \text{Text}, \text{Audio} \rangle$**

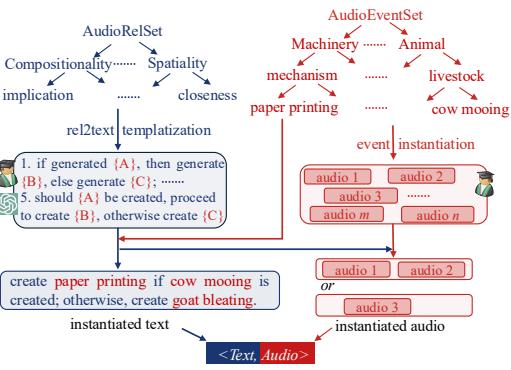
282 With the constructed audio event corpus in
 283 Sec. 3.1 and relation corpus in Sec. 3.2, we can further construct relation aware
 284 $\langle \text{Text}, \text{audio} \rangle$ pairs. Specifically, as is
 285 shown in Fig. 4, we first associate each of the
 286 100 relations in the relation corpus with metic-
 287 ulously curated 5 text description templates. We
 288 either manually write or query GPT-4o to gen-
 289 erate 5 text prompt templates precisely describ-
 290 ing the relation and accommodating the large
 291 language usage variation (see Fig. 4 line 4-8).
 292 Each template contains audio events name place-
 293 holder, we instantiate the template by replacing
 294 the placeholder with real audio event name to
 295 obtain the text prompt. To accommodate the
 296 synonymy of audio event name, we maintain a
 297 synonym list for each audio event name, and ran-
 298 domly select one each time when instantiating
 299 the template. For example, the audio event name
 300 “hammer nailing” can be synonymously replaced
 301 by one of *[hitting, slapping, smacking, punching]*.
 302

303 To accurately describe the audio event with text, we adopt the “Head-Modifier Structure with
 304 Progressive Verb Form” approach. In this approach, the description begins with the subject or entity
 305 producing the audio (*e.g.*, “food”) as the head, emphasizing the primary source of the sound. The
 306 action is then specified using its present participle form (*e.g.*, “frying”) as the modifier to convey a
 307 sense of immediacy and highlight that the audio event is ongoing. For instance, instead of describing
 308 a sound as “frying food” or “fry food” it is labeled as “food frying audio,” where the subject (“food”)
 309 is foregrounded, and the action (“frying”) contextualizes the nature of the audio. This approach
 310 ensures clarity, aligns with the temporal context of the audio, and effectively captures the dynamic
 311 nature of the event. With the same audio events name, we can retrieve its relevant audio waveform
 312 data and generate the audio by following the relation (He et al., 2025).
 313

314 **4 EXPERIMENT**

315 **4.1 DATASET CONSTRUCTION**

316 Following the common setup in existing TTA model, the created audio is 10 second long with
 317 sampling rate 16 kHz. Based on the data creation method introduced in Sec. 3.3, in the training phase,
 318 we randomly construct 360 $\langle \text{Text}, \text{audio} \rangle$ pairs for each relation, and in total we have created
 319 36,000 pairs. In testing phase, we randomly construct 100 $\langle \text{Text}, \text{audio} \rangle$ pairs for each relation,
 320 ensuring no constructed pairs appear in the training dataset. As we follow the prior TTA models
 321 setting to create the audio to be 10 seconds long with sampling rate 16 kHz, the training audio dataset
 322 is 100 hours, and testing audio dataset is 28 hours. Since we decouple relation from audio events
 323 during dataset construction and the texts in training dataset are different from the texts in the testing



324 **Figure 4:** $\langle \text{Text}, \text{audio} \rangle$ pair generation illustration,
 325 which can generate nearly unlimited pairs
 326 with high diversity.

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

360

361

362

363

364

365

366

367

368

369

370

371

372

373

374

375

376

377

378

379

380

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

398

399

400

401

402

403

404

405

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

450

451

452

453

454

455

456

457

458

459

460

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

483

484

485

486

487

488

489

490

491

492

493

494

495

496

497

498

499

500

501

502

503

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540

541

542

543

544

545

546

547

548

549

550

551

552

553

554

555

556

557

558

559

560

561

562

563

564

565

566

567

568

569

570

324 dataset, the constructed training and testing `<text, audio>` pairs have no overlap and differ from
 325 each other significantly.
 326

327 **4.2 EVALUATION METRIC**
 328

329 We accommodate both classic general evaluation metrics and relation aware evaluation metrics. For
 330 general evaluation, we follow traditional TTA works (Liu et al., 2024; 2023; Ghosal et al., 2023;
 331 Majumder et al., 2024) and adopt three metrics: Fréchet Audio Distance (FAD), Fréchet Distance
 332 (FD) (Heusel et al., 2017), Kullback–Leibler (KL) divergence. These three metrics measure the
 333 overall similarity in embedding space between reference audio and generated target audio without
 334 explicitly taking relation into account. Specifically, following the practice in prior TTA works,
 335 we extract the embeddings from VGGish (Hershey et al., 2017) model for FAD and KL metrics,
 336 embeddings from PANNs (Kong et al., 2020) model for FD metric.
 337

338 For relation aware evaluation, we adopt the multi-stage relation aware (*MSR*) evaluation protocol
 339 introduced in RiTTA (He et al., 2025). In *MSR* protocol, we first explicitly extract out audio events
 340 and relations (E', R') from generated audio, then further compare them with reference audio events
 341 and relations (E, R). To reflect if the model has generated but generated only the designated audio
 342 events and relations, *MSR* adopts *Presence*, *Relation correctness* and *Parsimony* score to gauge the
 343 quality of generated audio from different perspectives. Specifically, we report *mAPre*, *mARel* and
 344 *mAPar* scores for either separate relations or across all relations. More detailed information about
 345 *MSR* metric refers to RiTTA (He et al., 2025). To extract out audio event from generated audio, we
 346 finetune an audio event detection and tagging model on top of the pre-trained PANNs (Kong et al.,
 347 2020) model with 1 million training dataset. The mAP on 100,000 testing dataset achieves 0.91 for
 348 audio event detection, ensuring the finetuned model can extract out all potential audio events with
 349 high precision. To classify acoustic effect, we train another 7 acoustic effects classification model
 350 on top of the pre-trained PANNs model with 1 million training dataset. The accuracy rate on 100 k
 351 testing dataset achieves 95%.

352 **4.3 BENCHMARKING METHODS**
 353

354 We exhaustively benchmark 9 most recent general TTA models: AudioLDM (Liu et al., 2023),
 355 AudioLDM 2 (Liu et al., 2024), MakeAnAudio (Huang et al., 2023), AudioGen (Kreuk et al.,
 356 2023), Tango (Ghosal et al., 2023), Tango 2 Majumder et al. (2024), LAFMA (Guan et al., 2024),
 357 Affusion (Xue et al., 2024) and TangoFlux (Hung et al., 2024). They are pretrained on general TTA
 358 dataset (Gemmeke et al., 2017; Kim et al., 2019). For benchmarking, we choose their released model
 359 to generate a 10 second audio from the text prompt, detailed configuration is in Table III in Appendix.
 360

361 We further benchmark two agentic workflow based methods, in which we leverage open-sourced
 362 Qwen family LLM acting as an agent to analyze the input text and output the separate audio events
 363 an TTA model needs to generate. At the same time, the same LLM works as the third agent to output
 364 the python code that merges the audios generated by the TTA model. The reason of experimenting
 365 agentic flow is to see if we can decompose the relation aware generation task into simple single audio
 366 event generation task. Detailed implementation of our agentic workflow is provided in Appendix .1.
 367

368 **4.4 BENCHMARKING RESULT ON EXISTING TTA MODELS**
 369

370 The benchmarking result is given in Table 2, from which we can observe that all existing TTA models
 371 perform poorly on relation aware TTA generation. Similar to RiTTA (He et al., 2025), we also find the
 372 contradictory evaluation result between general evaluation and relation aware evaluation, which shows
 373 the speciality of relation aware TTA task. Among all the benchmarking methods, AudioGen (Kreuk
 374 et al., 2023) and TangoFlux (Hung et al., 2024) perform the best. While AudioGen (Kreuk et al.,
 375 2023) achieving the best in *mAPar* (relation parsimony) and *mAMSR*, TangoFlux (Hung et al., 2024)
 376 stays the best-performing in *mAPre* and *mARel* which mean it excels at accurately generating the
 377 target audio events and corresponding relation. However, almost all benchmarking methods achieves
 378 less than 10% percent accuracy rate across all relation aware evaluation metrics, which in turn verify
 379 the necessity to introduce new large-scale benchmark tailored for relation aware TTA research.
 380

381 Furthermore, both the two comparing agentic flow baselines perform poorly, they perform substan-
 382 tially worse than most existing existing TTA approaches. This poor performance highlights a critical
 383

378 Table 2: Quantitative benchmarking result on our introduced benchmark. mAPre, mARel and mAPar
 379 are in 10^{-2} . mAPre and mARel can be treated as *presence, relation correctness* percentage ratio,
 380 they lie in range $[0, 100]$. mAPar score also lies within $[0, 100]$. mAMSR (%) lies in range $[0, 1]$

Eval Way	Model	#Param	General Evaluation			Relation Aware Evaluation %(\uparrow)			
			FAD \downarrow	KL \downarrow	FD \downarrow	mAPre	mARel	mAPar	mAMSR
Zero-Shot	AudioLDM (s-full) 2023	185 M	4.02	21.23	22.36	3.47	0.91	2.95	0.73
	AudioLDM (l-full) 2023	739 M	4.13	22.05	23.03	3.10	0.79	2.63	0.63
	AudioLDM 2 (l-full) 2024	844 M	4.54	22.90	30.53	0.35	0.04	0.31	0.03
	MakeAnAudio 2023	452 M	5.10	50.97	30.49	4.75	0.88	4.05	0.73
	AudioGen 2023	1.5 B	7.97	25.19	32.29	11.3	2.84	9.13	2.22
	LAFMA 2024	272 M	25.85	269.54	65.27	0.96	0.15	0.45	0.07
	Affusion 2024	1.1 B	4.13	42.59	31.17	6.71	1.41	4.07	0.79
	Tango 2023	866 M	7.47	64.10	28.28	4.46	0.98	3.67	0.79
	Tango 2 2024	866 M	9.59	65.24	35.50	9.68	2.48	5.49	1.29
	TangoFlux 2024	576 M	6.01	26.73	30.00	12.38	3.34	7.28	1.77
Agen tic	Qwen2 7B+TangoFlux	-	9.98	142.87	39.20	3.53	0.77	2.25	0.04
	Qwen2.5 32B+TangoFlux	-	9.70	140.56	38.65	3.79	0.96	2.41	0.60

394 limitation: simply scaling up current TTA methods without fundamentally enhancing their relation
 395 aware modeling capability is unlikely to succeed. In this light, the benchmark introduced in this paper
 396 is not merely a comparison tool but a catalyst—providing the necessary structure, evaluation, and
 397 motivation to drive genuine advances in relation aware TTA research.

4.5 TWO INTUITIVE WAYS TO IMPROVE RELATION AWARE MODELING

400 Table 3: Quantitative result comparison on testset between finetuning (ft) and training from
 401 scratch (scratch) on curated 100 hours dataset.

Train Way	Model	#Param	General Evaluation			Relation Aware Evaluation %(\uparrow)			
			FAD \downarrow	KL \downarrow	FD \downarrow	mAPre	mARel	mAPar	mAMSR
ft	Tango 2023	866 M	3.88	33.26	21.30	14.58	4.18	10.16	2.73
	Tango 2 2024	866 M	4.06	22.39	20.32	15.53	4.63	10.21	2.86
	TangoFlux 2024	576 M	1.29	9.68	16.44	28.57	8.02	20.84	5.58
scratch	Tango 2023	866 M	3.63	22.34	20.16	14.89	3.69	10.98	2.64
	TangoFlux 2024	576 M	1.64	17.82	11.72	16.68	3.82	12.01	2.58

411 Two intuitive strategies to enhance relation aware modeling in existing TTA methods are (i) finetuning on our
 412 curated dataset and (ii) training from scratch. This dual perspective not only tests the feasibility of our benchmark but
 413 also evaluates the potential of transferring general TTA domain knowledge into relation aware settings. To this end,
 414 we apply both training strategies to three representative
 415 baselines: Tango (Ghosal et al., 2023), Tango 2 (Majumder
 416 et al., 2024), and TangoFlux (Hung et al., 2024). The re-
 417 sults in Table 3 reveal a clear trend: both finetuning and
 418 training from scratch substantially improve relation aware
 419 performance, validating the effectiveness of our bench-
 420 mark as a testing ground for relation aware TTA. Notably,
 421 TangoFlux benefits the most from finetuning, indicating
 422 that cross-domain TTA knowledge can be effectively trans-
 423 ferred to relation aware tasks. In contrast, Tango shows
 424 little difference between the two strategies, suggesting that
 425 model architecture and inductive bias may affect the extent
 426 to which general TTA knowledge can be leveraged. These findings highlight our benchmark’s unique
 427 role in uncovering such model-specific behaviors and point to an open research direction: how to
 428 best exploit general TTA knowledge to scale up relation aware TTA, and conversely, how relation
 429 aware training can reciprocate general TTA advances. We visualize the generated audio comparison
 430 between TangoFlux in zero-shot and finetuned base inference mode in Fig. 5, from this figure we can
 431 clearly see that finetuning on our curated dataset benefits relation aware modeling.

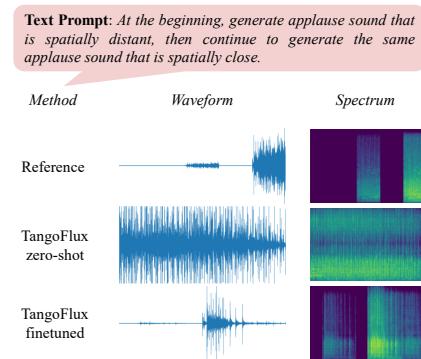


Figure 5: Qualitative comparison be-
 429 tween zero-shot and finetune based TangoFlux inference on one text prompt.

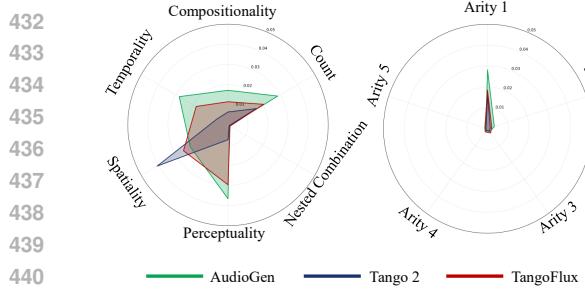


Figure 7: mAMSR regarding 6 main relation category and 5 relation Arity in Zero-shot setting.

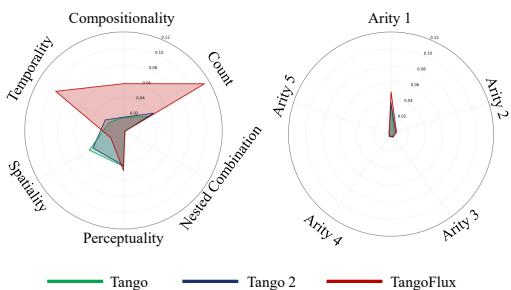


Figure 8: mAMSR regarding 6 main relation category and 5 relation Arity in finetuning setting.

To further investigate the role of training datasize, we extend both finetuning and training from scratch experiments to larger datasets of 200 hours and 300 hours. As is shown in Fig. 6, the mAMSR trend reveals two distinct behaviors: finetuning yields strong early gains but quickly saturates when the datasize approaches 300 hours, whereas training from scratch continues to improve substantially with increasing data. This divergence underscores an important insights: scaling relation aware TTA models ultimately requires massive datasets, and reliance on finetuning alone may be insufficient for long-term progress. Our benchmark is therefore essential: it not only provides the controlled scaling environment needed to expose these trends, but also offers the first practical platform to systematically study how training strategy and datasize interact in advancing relation aware TTA.

We visualize mAMSR regarding 5 main relations and relation Arity of three strong methods in Zero-shot setting (Fig. 7) and finetuning based setting (Fig. 8). From these two figures, we can observe that finetuning can improve the relation aware modeling capability in general. In zero-shot setting, AudioGen (Kreuk et al., 2023) performs better than the other two on main categories including *Temporality*, *Count* and *Perceptuality*. While in finetuning setting, TangoFlux (Hung et al., 2024) becomes the best-performing method. However, all methods in both setting perform terrible in *Nested Combination* or Arity larger than 1. Our proposed benchmark enables researchers to tackle these challenges at scale.

4.6 MORE INVESTIGATION ON EXISTING TTA MODEL ANALYSIS

Relation aware TTA demands not only the correct presence of target audio events but also the faithful preservation of their underlying relations. However, current TTA methods (Hung et al., 2024; Ghosal et al., 2023; Xue et al., 2024) remain narrowly focused on single-event generation, leaving them ill-equipped to handle multi-event, relation aware prompts. Table 4 makes this gap explicit: while TangoFlux (Hung et al., 2024), the state-of-the-art general TTA model, achieves 75% accuracy on single-event prompts, its performance collapses to just 12% for multi-event correctness and a mere 3% for relation fidelity. This dramatic degradation exposes a fundamental blind spot in existing approaches—relation aware modeling is virtually unaddressed. Our benchmark directly targets this deficiency, offering the first systematic platform to quantify and dissect these failures. By doing so, it not only diagnoses the shortcomings of current TTA methods but also establishes the essential foundation for driving genuine advances in relation aware TTA.

Conclusion. By curating AudioEventSet (110 events) and AudioRelSet (100 relations) with a systematic pair-generation strategy, Aurelius enables large-scale benchmarking and analysis. Our evaluations expose the sharp limitations of existing TTA models in relation fidelity. Aurelius establishes the first comprehensive benchmark for relation aware TTA and paves the way for future research on modeling complex event–relation dynamics in sound.

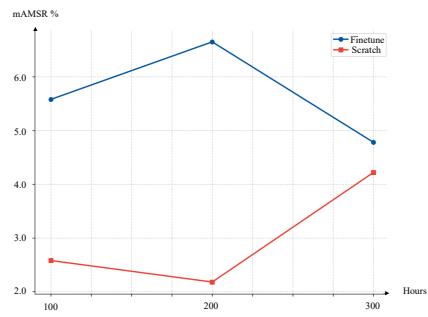


Figure 6: mAMSR variation w.r.t. training datasize (100 h, 200 h and 300 h).

Table 4: Audio event and relation accuracy of TangoFlux generation under different setting.

Description	Accu.
Event (single event, no relation)	75%
Event (multi-event, relation aware)	12%
Relation (multi-event, relation aware)	3%

486 REFERENCES
487

488 Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon,
489 Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. MusicLM: Generating
490 Music from Text. In *arXiv preprint arXiv:2301.11325*, 2023. 2

491 Peihao Chen, Yang Zhang, Mingkui Tan, Hongdong Xiao, Deng Huang, and Chuang Gan. Generating
492 Visually Aligned Sound from Videos. *IEEE Transactions on Image Processing (TIP)*, 2020. 2

493

494 Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexan-
495 dre Défossez. Simple and Controllable Music Generation. In *Advances in Neural Information
496 Processing Systems (NeurIPS)*, 2023. 2

497 Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: an Open
498 Dataset of Human-labeled Sound Events. *IEEE/ACM Transactions on Audio, Speech, and Lan-
499 guage Processing (TASLP)*, 30:829–852, 2022. 1, 3, 4

500

501 J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and
502 M. Ritter. Audio Set: An Ontology and Human-Labeled Dataset for Audio Events. In *IEEE
503 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. 1, 3, 7

504

505 Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-Audio
506 Generation using Instruction Tuned LLM and Latent Diffusion Model. In *ACM International
507 Conference on Multimedia (ACMMM)*, 2023. 1, 2, 3, 7, 8, 9, 17

508

509 Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Reddy Evuru, Ra-
510 maneswaran S, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. CompA:
511 Addressing the Gap in Compositional Reasoning in Audio-Language Models. In *International
512 Conference on Learning Representations (ICLR)*, 2024. 2, 4

513

514 Wenhao Guan, Kaidi Wang, Wangjin Zhou, Yang Wang, Feng Deng, Hui Wang, Lin Li, Qingyang
515 Hong, and Yong Qin. LAFMA: A Latent Flow Matching Model for Text-to-Audio Generation. In
516 *Interspeech*, 2024. 1, 3, 7, 8, 17

517

518 Yuhang He, Yash Jain, Xubo Liu, Andrew Markham, and Vibhav Vineet. RiTTA: Modeling Event
519 Relations in Text-to-Audio Generation. In *Conference on Empirical Methods in Natural Language
520 Processing (EMNLP)*, 2025. 1, 2, 3, 4, 6, 7, 16

521

522 Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing
523 Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss,
524 and Kevin Wilson. CNN Architectures for Large-Scale Audio Classification. In *International
525 Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017. 7

526

527 Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter.
528 GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In
529 *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 7

530

531 IJ Hirsh, C Milliman, and F Darley. Brain Mechanisms Underlying Speech and Language, 1967. 1

532

533 Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. *Advances in
534 Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2

535

536 Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin
537 Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio Generation with Prompt-enhanced
538 Diffusion Models. *International Conference on Machine Learning (ICML)*, 2023. 7, 8, 17

539

540 Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Rafael Valle, Bryan Catanzaro,
541 and Soujanya Poria. TangoFlux: Super Fast and Faithful Text to Audio Generation with Flow
542 Matching and Clap-Ranked Preference Optimization, 2024. 1, 3, 7, 8, 9, 14, 17

543

544 Vladimir Iashin and Esa Rahtu. Taming Visually Guided Sound Generation. In *British Machine
545 Vision Conference (BMVC)*, 2021. 2

540 Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating
 541 captions for audios in the wild. In *Conference of the North American Chapter of the Association*
 542 *for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2019. 1, 3, 7
 543

544 Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark Plumbley. PANNS:
 545 Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. In *IEEE/ACM*
 546 *Transactions on Audio, Speech, and Language Processing (TASLP)*, 2020. 7

547 Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet,
 548 Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually Guided Audio Generation.
 549 *International Conference on Learning Representations (ICLR)*, 2023. 2, 3, 7, 8, 9, 17
 550

551 Brenden M. Lake, Ruslan Salakhutdinov, and Joshua B. Tenenbaum. Human-level Concept Learning
 552 through Probabilistic Program Induction. *Science*, 350(6266):1332–1338, 2015. 1
 553

554 Huan Liao, Haonan Han, Kai Yang, Tianjiao Du, Rui Yang, Zunnan Xu, Qinmei Xu, Jingquan
 555 Liu, Jiasheng Lu, and Xiu Li. BATON: Aligning Text-to-Audio Model with Human Preference
 556 Feedback. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2024. 2, 3
 557

558 Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow Match-
 559 ing for Generative Modeling. In *International Conference on Learning Representations (ICLR)*,
 560 2023. 1, 2

561 Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D
 562 Plumbley. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. *International*
 563 *Conference on Machine Learning (ICML)*, 2023. 2, 3, 7, 8, 17

564 Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu
 565 Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning Holistic Audio Generation
 566 With Self-Supervised Pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language*
 567 *Processing (TASLP)*, 2024. 1, 2, 3, 7, 8

568

569 Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional Vi-
 570 sual Generation with Composable Diffusion Models. In *European Conference on Computer*
 571 *Vision (ECCV)*, 2022. 2

572 Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya
 573 Poria. Tango 2: Aligning Diffusion-based Text-to-Audio Generations Through Direct Preference
 574 Optimization. In *ACM International Conference on Multimedia (ACMMM)*, 2024. 7, 8, 17
 575

576 Roy Sheffer and Yossi Adi. I Hear Your True Colors: Image Guided Audio Generation. In *IEEE*
 577 *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. 2

578 Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based Generative Modeling in Latent Space. In
 579 *Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2
 580

581 Somin Wadhwa, Silvio Amir, and Byron C. Wallace. Revisiting relation extraction in the era of large
 582 language models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*,
 583 2023. 2

584

585 Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyu Wu. AudioTime: A Temporally-aligned Audio-
 586 Text Benchmark Dataset. *IEEE International Conference on Acoustics, Speech and Signal Pro-
 587 cessing (ICASSP)*, 2025a. 2, 3, 4

588

589 Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyu Wu. PicoAudio: Enabling Precise Temporal
 590 Controllability in Text-to-Audio Generation. In *IEEE International Conference on Acoustics,
 591 Speech and Signal Processing (ICASSP)*, 2025b. 2

592 Jinlong Xue, Yayue Deng, Yingming Gao, and Ya Li. Auffusion: Leveraging the Power of Diffusion
 593 and Large Language Models for Text-to-Audio Generation. *IEEE/ACM Transactions on Audio,
 594 Speech, and Language Processing (TASLP)*, 2024. 1, 3, 7, 8, 9, 17

594 Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu.
595 Diffsound: Discrete Diffusion Model for Text-to-sound Generation. *IEEE Transactions on Audio,*
596 *Speech and Language Processing*, 2022. [2](#), [3](#)

597 Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and Jeremy R Reynolds. Event
598 Perception: A Mind-Brain Perspective, 2007. [1](#)

600 Aimen Zerroug, Mohit Vaishnav, Julien Colin, Sebastian Musslick, and Thomas Serre. A Bench-
601 mark for Compositional Visual Reasoning. In *Advances in Neural Information Processing Sys-*
602 *tems (NeurIPS)*, 2022. [2](#)

603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647

648 A APPENDIX
649650 .1 AGENTIC AUDIO SCENE GENERATION WORKFLOW
651652 We design an agentic workflow that integrates large language model (LLM) reasoning with state-of-
653 the-art text-to-audio (TTA) generation. The workflow operates in three stages:654 **1. Scene Planning** Given a natural-language description and a target duration, we guide a reasoning-
655 capable LLM (Qwen) with a structured system prompt to decompose the input into a JSON plan. The
656 prompt enforces JSON-only output, concise event descriptions in “head + present participle” style,
657 and relation aware scheduling rules. The full system prompt is provided below:
658659 **Audio Scene Planner Prompt**
660661 You are an **audio scene planner** for relation-aware text-to-audio (TTA) generation. Perform
662 all reasoning INTERNALLY and output ONLY valid JSON with the final scheduling results.
663 Never include explanations or chain-of-thought. Your JSON MUST be parsable and follow
664 the schema exactly.665 **OBJECTIVE**666 Given (1) a natural-language text prompt describing an audio scene with multiple events
667 and relations, and (2) a target total duration in seconds, decompose the scene into concise
668 sub-prompts (each corresponds to one independent audio event to be synthesized by a TTA
669 model), and schedule them on the global timeline with start times and durations.
670671 **OUTPUT CONTRACT (JSON ONLY)**672 {
673 "total_duration_sec": <float>,
674 // equals the requested total duration
675 "sub_prompts": [
676 {
677 "id": "E1", // short unique id
678 "text": "<short English sub-prompt>",
679 // head + present participle; may include light modifiers
680 "start_sec": <float>, // >= 0
681 "duration_sec": <float>
682 // > 0, and start_sec + duration_sec <= total_duration_sec
683 }
684 // ... typically <= items total
685]
686 }686 **LANGUAGE FOR SUB-PROMPTS**687

- Use concise English in “head + present participle” style: e.g., “door bell ringing
688 audio”, “footsteps running audio”.
- Add light, meaningful modifiers when clearly implied by text: “as background”,
689 “approaching”, “departing”, “slight reverberation”, “time-stretched”, “amplified”,
690 “attenuated”, “balanced against X”, “blended with Y”.
- Avoid redundant words; keep each sub-prompt single-sentence and ≤ 18 words.

694 **RELATIONS AND HOW TO REFLECT THEM (IMPLICITLY VIA SCHEDULING + WORDING)**695

- **Sequencing:** “then/after/next/first...then...” \rightarrow schedule sequentially with a small
696 gap ≈ 0.1 s.
- **Simultaneity / Mix / Background** \rightarrow allow overlaps; backgrounds can span large
697 portions of the timeline.
- **Approaching / Departing** \rightarrow keep in wording (“approaching”/“departing”). No
698 mandatory duration change.

- **OR / XOR** → choose the most natural option; do not include the unchosen one.
- **NOT / prohibit** → exclude that event entirely.
- **IF...THEN...ELSE** → choose the most sensible branch; output only the chosen branch.
- **Count** → if an explicit number of items/events is requested, match it.
- **Repetition** → instantiate repeated events (e.g., bell ringing three times) as multiple sub-segments or one sustained segment if implied.
- **Proximity / Closeness / Farness** → reflect via wording only (“distant thunder”, “near crowd”). No strict timing rules.

713 DURATION AND SCHEDULING RULES

- Respect total duration: sum of all segments should match the requested duration ($\pm 0.25s$); if off, adjust proportionally.
- Each segment `duration_sec` MUST be integer.
- Choose reasonable segment durations:
 - transient cues (e.g., bell, door knock, gunshot) → 1–2s
 - medium actions (e.g., footsteps, typing, sawing) → 2–6s
 - ambient backgrounds (e.g., rain, wind, crowd murmur) → long spans (often entire duration)
- `start_sec` ≥ 0 ; `end` \leq total duration.
- On conflicts, preserve explicit relations first; compress lightly but keep segments $\geq 0.5s$.
- Keep the number of sub-prompts concise (typically ≤ 5).

729 FORMATTING RULES

- Output VALID JSON only. No comments, no trailing commas, no text outside JSON.
- Floats may be given with 1–2 decimals.
- Ensure “`start_sec + duration_sec ≤ total_duration_sec`” for all segments.

736 **2. Segment Synthesis** Each sub-prompt is independently synthesized into audio using Tan-
 737 goFlux (Hung et al., 2024), a flow-matching-based TTA model. We enforce consistency between the
 738 planned duration and the generated waveform length.

739 **3. Timeline Stitching** All generated audio segments are placed on a global timeline according to
 740 their planned start times. Overlapping segments are blended with linear crossfades, and amplitude
 741 normalization is applied to prevent clipping. This yields a coherent audio scene that respects both
 742 semantic relations (e.g., sequencing, simultaneity) and global duration constraints.

744 .2 AUDIO EVENT CATEGORY CURATION DETAIL

746 .3 AUDIO EVENTS RELATION CORPUS

748 .4 BENCHMARKING MODEL INFERENCE SETTING

756
 757 Table I: *AudioEventSet* corpus detail. We list all 110 event classes, which are deriving from 7 main
 758 categories and 23 sub-categories.

Main Category	Sub-Category	Names	Description
761 762 Animal 763 (22) 764 765 766 767 768 769 770	wild ground animal	lion roaring, wolf howling, donkey braying, cricket chirping, frog croaking, horse neighing	771 live in the wild
	domestic animal	772 dog barking, cat meowing, dog growling, cat spurring	773 live in domestic setting
	774 livestock	775 pig oinking, sheep bleating, cow mooing, 776 rooster crowing, duck quacking	777 domesticated livestock
	778 wild animal	779 cuckoo calling, birds chorus, seagull cawing, peacock rattling, blue jay whistling, nightingale singing, fly buzzing	780 animals in the wild
781 Human 782 (21) 783 784 785 786 787 788 789 790 791 792	793 human voice	794 baby crying, laughing, shouting, 795 whistling, coughing, snoring, sneezing, chewing, 796 burping, farting	797 human use vocal tract
	798 human speech	799 male speech, female speech, child speech, group talk	800 speech audio
	801 hands action	802 finger snapping, clapping	803 audio by action
	804 group action	805 group clapping, cheering, group talking	806 audio by a group
	807 locomotion	808 running, footsteps	809 audio by movement
810 Machinery 811 (13) 812 813 814 815 816 817 818 819 820 821 822 823 824 825 826 827 828 829 830 831 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 9999	817 alarm	818 siren, door bell, car horn, 819 bicycle bell, telephone ringing, 820 telephone dialing, boat horn, 821 ratchet and pawl clicking, 822 camera shuttering, 823 printer printing, engine revving, 824 clock ticking, paper shredding	825 machinery alarming
	826 mechanism	827	828 mechanism audio
829 Human-Obj 830 Interaction 831 (18) 832 833 834 835 836 837 838 839 840 841 842 843 844 845 846 847 848 849 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 9999	830 tools	831 hammer nailing, wood sawing, 832 pen writing, wood chopping, rasping 833 dish audio, silverware audio, 834 food frying, vegetable chopping 835 toilet flushing, pouring water, 836 keyboard typing, door slamming, 837 cupboard open or close, 838 drawer open or close, packing tape, 839 dentist drilling, door knocking	840 human use tools
	841 culinary	842	843 in kitchen setting
	844 work	845	846 audio during work
847 Obj-Obj 848 Interaction 849 (15) 850 851 852 853 854 855 856 857 858 859 860 861 862 863 864 865 866 867 868 869 870 871 872 873 874 875 876 877 878 879 880 881 882 883 884 885 886 887 888 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 910 911 912 913 914 915 916 917 918 919 920 921 922 923 924 925 926 927 928 929 930 931 932 933 934 935 936 937 938 939 940 941 942 943 944 945 946 947 948 949 950 951 952 953 954 955 956 957 958 959 960 961 962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 990 991 992 993 994 995 996 997 998 999 9999	850 impact audio	851 key jingling, ball bouncing, 852 pen clicking, wind chime, 853 car emergency braking	854 impact effect
	854 friction audio	855 knife sharpening, sandpaper scraping, 856 plastic scratching, string rubbing	857 friction effect
	857 dropping audio	858 coin dropping, glass clinking, 859 metal dropping	860 dropping effect
	859 explosion	860 gunshot, firework, artillery fire	861 explosion effect
862 Music 863 (11) 864 865 866 867 868 869 8699	869 music instrument	870 plucked string, piano keyboard, 871 bowed string, wind string, brass, 872 harmonica, accordion	873 musical instruments
	873 singing	874 female singing, male singing, 875 child singing, group singing	876 singing audio
877 Nature 878 (10) 879 880 881 882 883 884 885 886 887 888 889 8899	889 water	890 water bubbling, ocean wave, 891 water dripping, water flowing, 892 water boiling	893 water movement
	893 weather	894 thunder, wind, rain	895 nature weather
	895 nature change	896 wood cracking, rustling leaves	897 natural change

810
 811
 812
 813
 814 Table II: *AudioRelSet* corpus detail. We introduce 21 basic relations, and advanced 79 nested
 815 combination relations, resulting in a total of 100 relations – 9 times larger than the relation corpus
 816 proposed in RITTA (He et al., 2025). *AudioRelSet* maximumly covers all potential relations that
 817 audio events may exhibit in either the physical world or linguistic description. It is worth noting that
 818 *AudioRelSet* is open-ended. By nesting existing relations, we can potentially construct massive new
 819 relations.

820 Category	821 Relation Name	822 Explanation	823 Event Arity	824 Sample prompt
825 Temporality (4)	826 precedence	827 before	828 binary	829 audio {A} followed by {B}
	830 succession	831 after	832 binary	833 create audio {A} after {B}
	834 simultaneity	835 same time	836 binary	837 {A} and {B} simultaneously
	838 periodicity	839 cyclic	840 unary	841 create audio {A} periodically
842 Spatiality (5)	843 closeness	844 spatial close	845 binary	846 {A} is closer than {B}
	847 farness	848 spatial far	849 binary	850 {A} is farther than audio {B}
	851 proximity	852 equal-dist	853 binary	854 {A} and {B} the same dist
	855 approaching	856 moving close	857 unary	858 {A} is moving closer
	859 departing	860 moving away	861 unary	862 {A} is moving further away.
863 Count (1)	864 count	865 number	866 n-ary	867 3 audios: {A}, {B} and {C}
868 Perceptuality (6)	869 balancing	870 level balance	871 binary	872 {A} dominates, {B} fades
	873 blending	874 mix audios	875 binary	876 {A} and {B} are mixed
	877 reverberation	878 reverberant	879 unary	880 generate audio {A} in canyon
	881 time-stretching	882 speed manipulate	883 unary	884 stretch audio {A} in time scale
	885 amplification	886 become louder	887 unary	888 amplify {A} to be louder
	889 attenuation	890 less loudly	891 unary	892 attenuate {A} to be quieter
893 Composition- ality (5)	894 conjunction	895 logical AND	896 binary	897 create both {A} and {B}
	898 disjunction	899 logical OR	900 binary	901 create {A} or {B}, or both
	902 negation	903 logical NOT	904 unary	905 do not generate audio {A}
	906 exclusive-or	907 logical XOR	908 binary	909 generate {A} or {B}, not both
	910 implication	911 if-then-else	912 ternary	913 if {A}, then {B}, else just {C}
893 Nested Combination (79)	914 Temp + Spat (4)	915 Temp + Spat	916 binary	917 {A} before approaching {B}
	918 Temp + Percep (8)	919 Temp + Percep		920 reverb. {A}, succeeded by {B}
	921 Percep + Comp (12)	922 Percep + Comp		923 stretched {A} or {B}, not both
	924 Spat + Comp (4)	925 Spat + Comp		926 approaching {A} or {B}, not both
	927 Temp + Comp (6)	928 Temp + Comp	929 ternary	930 {A} first, then {B} or {C}
	931 Percep + Comp (1)	932 Percep + Comp		933 mix {A} with {B}, or {C}
	934 Comp + Comp (1)	935 Comp + Comp		936 {A} and {B}, or {A} and {C}
	937 Spat + Comp (5)	938 Comp + Comp		939 {A} and {B}, or {A} and {C}
	940 Spat + Comp + Percep (2)	941 Comp + Comp	942 quaternary	943 {A} and {B}, or {A} and {C}
	943 Temp + Comp (4)	944 Temp + Comp		945 audio {A} or {B} first, followed by {C} or {D}
	946 Comp + Comp (7)	947 Comp + Comp		948 {A} or {B} first, then {C} or {D}
	949 Temp + Comp (3)	950 Temp + Comp		951 {A} before {B} first, then {C}
	952 Spat + Comp (9)	953 Spat + Comp	954 quinary	954 before {D} or {E}
	955 Comp + Comp (9)	956 Comp + Comp		957 if {A} closer than {B}, then {C} closer than {D}, else {E}
	958 Count + Comp (4)	959 Count + Comp		960 if {A} and {B}, then {C} and {D} else {E}
				961 if {A}, {B}, {C}, then {D}, else {E}

864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884

885

Methods	Setting
AudioLDM (S-Full) (2023)	guidance_scale=5, random_seed=42, n_candidates=3
AudioLDM (L-Full) (2023)	guidance_scale=5, random_seed=42, n_candidates=3
AudioLDM 2 (L-Full) (2023)	guidance_scale=3.5, random_seed=45, n_candidates=3
MakeAnAudio (2023)	ddim_steps = 100, scale = 3.0
AudioGen (2023)	model name: audiogen-medium
Auffusion (2024)	num_steps = 100, guidance=7.5, num_samples=1
LAFMA (2024)	num_steps = 200, guidance=3, num_samples=1
Tango (2023)	num_steps = 200, guidance=3, num_samples=1
Tango 2 (2024)	num_steps = 200, guidance=3, num_samples=1
TangoFlux (2024)	num_steps = 50, guidance=3, num_samples=1

896 Table III: Detail setting for each TTA method.
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917