GASA: A STRUCTURED ALIGNMENT BENCHMARK FOR TEXT-TO-AUDIO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-audio generation has advanced faster than the methods used to evaluate it. Models are known to miss parts of a prompt, but prevailing measures judge a generated clip against its prompt as a whole and cannot show which parts are missed or where research should focus. We introduce the Generative Audio Structured Alignment (GASA) benchmark. GASA treats prompt fidelity as structured alignment, organizing the properties of a scene description into 25 fields and judging each field with a large audio-language model. Because such judges may answer from text-only priors without listening, we validate judge probes on real caption-audio pairs against a Gaussian-noise control, keeping for each field the probe whose accuracy depends most on the audio. Human raters answering the same probes on generated clips confirm that the judge detects 23 of the 25 fields in generated audio and set the ceiling each field score can attain. Using this validated judge, we evaluate 11 models on 2,500 generation prompts (100 per field). Scores vary far more across fields than across models, and every model leaves several fields near the control. Source identity is strongest, static acoustic properties weakest, and pitch and timbre are rendered more faithfully as changes than as static levels. GASA pinpoints where each model fails, giving researchers specific targets and a measure of progress toward models that realize every part of a prompt. `https://gasa-demo.github.io/`

1 INTRODUCTION

Text-to-audio (TTA) generation has progressed rapidly, and recent models can synthesize realistic and semantically rich sound scenes from free-form natural-language prompts. The central question has shifted from whether a generated clip sounds plausible to whether it fully realizes the content the prompt specifies. Evaluation has advanced more slowly, and the most widely used measures judge a clip globally, comparing it against a reference distribution or scoring its overall similarity to the prompt. Even recent audio question-answering and captioning metrics, more sensitive to fine-grained meaning, assess the whole response as a single unit.

Assigning a single score to a generated audio clip does not evaluate which aspects of a prompt a generated clip realizes and which it misses. A clip may contain the sound events named in a prompt, but in the wrong order or with the wrong acoustic character. A global comparison may still rate it highly, since these measures primarily reward the presence of the dominant events. Similarity-based metrics are generally insensitive to such fine-grained differences, unable to separate inverted or attribute-swapped descriptions, e.g., “a dog barking before thunder” vs. “thunder before a dog barking” (Ghosh et al., 2024). To diagnose prompt alignment at a finer grain, we move the unit of evaluation from the whole prompt to the individual properties it expresses.

We treat this as a problem of structured alignment. To check a single property in an audio clip, we use an audio-language model as the judge: given the clip and a natural-language question about one property, it answers whether the clip supports that property. The same model can answer questions about objects, attributes, timing, and relations between sounds, so one judge covers every property.

This convenience comes with a risk. Audio-language models may lack such capabilities and may also carry strong text-only priors, enabling them to often answer correctly without attending to the audio. Recent work finds they keep much of their accuracy even when the real clip is replaced by

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

noise (Zang et al., 2025). A judge with this weakness cannot be trusted to score generated audio, because a low score might mean the generator failed or might mean the judge never listened. Showing that the judge actually listens and detects such fine-grained properties is therefore a precondition for property-level evaluation.

We address this in three parts: 1) We organize what an audio scene description can express into a hierarchy of 25 fields. 2) We build probes from real caption-audio pairs, where a human-written caption tells us which field values the clip supports, and we score each probe on the real clip and on a Gaussian-noise control. We compare three audio-language judges this way and select the one whose accuracy depends most on the audio across fields. For each field, we then keep the probe format whose accuracy depends most on the audio. 3) Using this validated judge and probe set as the measurement instrument, we introduce the Generative Audio Structured Alignment (GASA) benchmark that scores how well TTA models realize each field of a prompt. We focus on everyday sound scenes and do not cover speech or music.

Applying GASA to eleven TTA models yields a consistent picture: performance varies far more across fields than across models, with even the two strongest models statistically indistinguishable. Across models, sound sources and the dynamics of pitch, timbre, and distance are rendered well, while static acoustic attributes are barely above the control.

2 BACKGROUND

We review how text-to-audio alignment is currently measured, how visual and audio benchmarks have decomposed alignment into separately tested factors, and what is known about the reliability of audio-language models as judges.

Human listening tests are the gold standard for perceptual quality (Liu et al., 2023; Kreuk et al., 2023) but high costs limit their scale and frequency. Furthermore, existing protocols rate text–audio alignment holistically or along a few coarse aspects such as event presence and order (Liao et al., 2024; Kanamori et al., 2025). Automatic measures scale better but also share this holistic-scoring limitation. Distribution-level metrics such as Fréchet Audio Distance (Kilgour et al., 2019) and Kernel Audio Distance (Chung et al., 2025) assess audio quality but not prompt fidelity. Per-clip embedding measures such as CLAPScore (Elizalde et al., 2023) evaluate correspondance, but Ghosh et al. (2024) report that CLAP-based scoring performs near chance when distinguishing event order or attribute–source bindings; so CLAP similarity measures broad relevance but not compositional structure.

A more recent approach reformulates text-to-audio alignment as question answering, a route first developed for image generation (Hu et al., 2023; Lin et al., 2024). AQAScore (Kuan et al., 2026a) poses a targeted question about a clip to an audio-language model and reads the probability the model assigns to a 'yes' answer. It is more sensitive than embedding similarity scoring and agrees more closely with human judgment, but it still scores the whole description at once.

A separate judgment for each property instead requires decomposing alignment into individual factors, testing each under a controlled comparison, and reporting a per-factor profile instead of a single score. Such decomposition is more developed in the visual than audio domain. The controlled-pair recipe originates in understanding, where Winoground (Thrush et al., 2022) credits a model only when it matches both members of a caption pair built from the same words in a different order. In generation, COCO-Counterfactuals (Le et al., 2023) pairs images whose captions differ by a single edited element; T2I-CompBench (Huang et al., 2023) splits compositional generation into attribute binding, object relationships, and complex compositions; and T2V-CompBench (Sun et al., 2025) applies the approach to text-to-video.

Audio evaluation has followed a similar but more limited trajectory. PhyAVBench (Xie et al., 2025) builds paired prompts that differ in one physical variable and checks whether the generated sound changes as the physics would predict, but it stays within physical parameters in a text-to-audio-video setting. TTA-Bench (Wang et al., 2026) covers seven dimensions from quality to robustness and toxicity, but its one accuracy dimension asks whether the audio reflects the prompt as a whole. CompA (Ghosh et al., 2024) tests whether an audio-language model can perceive event order and attribute binding, but not whether a generator produced them. ELSA (Suzuki et al., 2026) comes closest to a per-property judgment for TTA models. It splits a prompt into its sound events with a

Table 1: 25-field, 7-dimension hierarchy.

Dimension	Fields
Source identity	object, action, material, count, human specifics
Acoustic attributes	pitch, loudness, timbre
Temporal properties	duration, repetition
Spatial properties	distance, direction
Dynamic properties	loudness dynamics, pitch dynamics, timbre dynamics, distance dynamics, direction dynamics, repetition dynamics
Global attributes	acoustic environment, location, emotion, style
Inter-source relations	temporal sequence, temporal concurrency, foreground-background

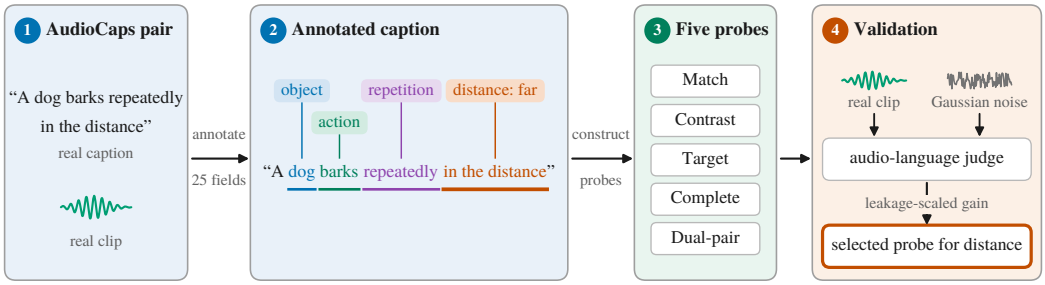

Figure 1: Pipeline for constructing and selecting a probe to evaluate a prompt on the distance field.

language model, separates the audio sources belonging to each event, scores each event against its phrase in a CLAP embedding space, and reports closer agreement with human ratings than whole-clip similarity. However, its evaluation unit is the event rather than the property: it scores whether a dog's bark is present, not separately whether the bark was loud or nearby, and it does not model event order. Existing benchmarks each address part of the problem, but none decomposes generated audio into the full range of properties a scene can express.

3 Structuring Sound Scene Descriptions into Fields

A natural-language description of a sound scene expresses more than which sound events occur. "A dog barks repeatedly in the distance" names a source and its action, but it also communicates how the barking unfolds in time and where the source sits relative to the listener. Thus, whether a clip matches this description is not one fact but several, one for each property. We organize the properties a description can express into a hierarchy of 25 fields grouped into seven dimensions (Table 1; definitions in Appendix A). These fields are grounded in caption corpus analysis (Appendix B), the AudioSet ontology (Gemmeke et al., 2017), and environmental sound perception research (Gaver, 1993; Guastavino, 2018).

To test these properties individually, we decompose audio descriptions into structured representations pairing expressed fields with the supporting text. We use the same representation for both validation captions and benchmark prompts. Qwen2.5-72B-Instruct (Qwen Team, 2024) annotates descriptions against the 25 fields (Figure 1); rules are in Appendix B and require explicit textual quotes to extract a property. For example, "A dog barks repeatedly in the distance" yields four fields: an object (dog), an action (barking), a repetition pattern (repeated), and a distance (far). For most fields, the value is stated directly by a noun or an adjective, such as "dog" or "distant". An exception are fields for properties that unfold over time or relate two sounds, which instead surface through cue words: a change in loudness as "getting louder," the order of two events as "then," event overlap as "while," and the relative source prominence as terms like "in the background."

4 USING LALMS TO JUDGE FINE-GRAINED ALIGNMENT

To evaluate individual scene properties rather than whole captions, we employ large audio-language models (LALMs) as automated judges. Before using an LALM in our benchmark, we first need to establish that its judgments reflect audio listening rather than text-only priors.

4.1 CONSTRUCTING PROBES TO EVALUATE STRUCTURED PROMPTS

We draw caption-audio pairs from AudioCaps (Kim et al., 2019), and use the fields from the human-written captions as target answers. To evaluate whether the judge attends to a specific property rather than relying on prior text biases, we test one field at a time; Figure 1 illustrates an example. We formulate four field-level comparison formats alongside a whole-caption baseline (full prompt templates in Appendix D):

- ***Contrast*:** Pairs the supported caption with one that replaces the target by its opposite (e.g., "in the distance" vs. "nearby"), asking which better describes the clip. This isolates fields with clean antonyms, but cannot cover open-class fields like object or action.
- ***Target*:** Asks directly whether the clip exhibits the target property. Each positive question is paired with a negative counterfactual expecting "No" to avoid yes-biased answers.
- ***Complete*:** Pairs the supported caption with one that removes the target detail or adds an unstated detail, asking which is more accurate and complete. Add and remove variants are mixed equally to neutralize length bias. Supports fields that lack direct opposites.
- ***Dual-pair*:** Poses four fixed candidate captions across two real audio clips. The correct choice shifts with the clip, and the pair is credited only if both answers are right, preventing purely text-based shortcuts by construction.
- ***Match*:** Poses a whole-caption yes/no question pairing the clip with its own caption or an unrelated caption, serving as a non-decomposed reference baseline (Kuan et al., 2026a).

4.2 VALIDATING PROBES

A probe is informative only when its answers depend on the audio rather than priors or question wording (Zang et al., 2025). We test this by evaluating each probe on both the real audio and a Gaussian noise control, paired with the same caption. Above-chance accuracy on the control indicates answer-format regularities or prior event co-occurrences rather than listening.

Predictions are obtained from the first-token probability assigned to candidate options: Yes/No for binary probes and option letters for multiple-choice formats (Kuan et al., 2026a; Robinson & Wingate, 2023). Since probe formats differ in chance accuracy C ($1/2$ for two-option probes; $1/16$ for the *Dual-pair*, which requires two correct four-option choices), raw accuracy is not directly comparable across formats. We therefore evaluate each probe via *leakage-scaled gain*: Gain $\times$ (1 $-$ Leakage), where Gain $= \max(0, (R - B)/(1 - B))$ measures audio-grounded improvement over the Gaussian-noise baseline B, and Leakage $= \max(0, (B - C)/(1 - C))$ penalizes probes that exceed chance C without listening (R is accuracy on real audio).

To select an LALM judge and an optimal probe format for each field, we score each field-probe combination using three candidates: Qwen2.5-Omni (Xu et al., 2025a), Qwen3-Omni (Xu et al., 2025b), and AudioFlamingo-Next (Ghosh et al., 2026). Qwen3-Omni yielded the highest leakage-scaled gain when averaged across all field-probe combinations (0.517, vs. 0.295 for Qwen2.5-Omni and 0.333 for AudioFlamingo-Next) and supplies the top-performing probe for 24 of the 25 fields (full validation matrix in Appendix C). We therefore adopt Qwen3-Omni as the judge in the remaining benchmarks. For each field, we retain the probe format that achieves the highest leakage-scaled gain under Qwen3-Omni, while holding out *Match* as a whole-caption baseline.

Table 3 lists the selected probe for each field alongside its leakage-scaled gain. The three inter-source relation fields demonstrate most clearly the necessity of field-level probing. On temporal sequence, temporal concurrency, and foreground-background, holistic *Match* scores drop to 0.159, 0.019, and 0.108, whereas the selected field-level probes reach 0.645, 0.716, and 0.637, showing that targeted probes are necessary to reliably capture how multiple sounds relate in time or prominence.

Table 2: Spearman correlation with mean listener ratings on RELATE. † marks results reported in prior work. All other rows are computed by us on this split; ELSA is run with its released code.

	REL (n=1,311)	**IS** (n=1,020)	**OS** (n=1,110)
CLAPScore (LAION-CLAP)	.412	.381	.278
CLAPScore (MS-CLAP)	.306	.272	.169
Human-CLAP	.463	.457	.278
ELSA	.345	.336	.240
RELATE predictor†	.383	–	–
AQAScore†	.556	–	–
Ours (*Match* + field indicators)	.478	.410	.300
Ours (fields + field indicators)	.488	.433	.314
Ours (*Match* + fields + field indicators)	.519	.458	.316

4.3 Agreement with Human Ratings

To test alignment with human listeners, we evaluate the judge zero-shot on RELATE's test split (Kanamori et al., 2025) across overall relevance (REL), event presence (IS), and temporal order (OS). As shown in Table 2, the judge's raw *Match* probability correlates with listener ratings at .480 on relevance, .419 on event presence and .277 on temporal order, above the CLAPScore rows of Table 2 on relevance and presence. ELSA (Suzuki et al., 2026) reports .379 on relevance, but only for the generated clips of this split; on those clips our judge's whole-caption score reaches .469. The ELSA row of Table 2 is our run of its released code on the full split.

Beyond the whole-caption comparison, we test whether decomposing alignment into structured fields captures additional information. A random forest fit on RELATE's training split from the 23 detectable fields and field-mention indicators (see Section 5.2) matches whole-caption prediction on relevance ($\rho = .488$ vs. .478). Combining field scores with *Match* significantly improves predictions for both relevance (.519, paired gain $+.042$, 95% CI [+.021, +.063]) and event presence (.458, gain $+.048$, 95% CI [+.021, +.076]), with temporal sequence contributing most to order prediction.

5 Generative Audio Structured Alignment Benchmark

To evaluate whether audio generated by a TTA model realizes each field specified in a prompt, we introduce the **Generative Audio Structured Alignment (GASA)** benchmark. GASA consists of a controlled prompt suite: 2,500 prompts structured as 1,250 minimal pairs spanning the 25 fields. Each minimal pair keeps the acoustic scene identical except for the field under test. This allows us to isolate whether a generator faithfully conditions on that specific field. Instead of reducing model capability to a single numerical score, GASA produces a 25-field alignment profile for each model.

5.1 Prompt Construction

There are 100 prompts (50 minimal pairs) for each of the 25 fields, with three pairing strategies:

1. Opposing values (20 fields): Pairs opposite descriptors, such as "nearby" vs. "in the distance" for distance, or "loud" vs. "faint" for loudness.
2. Sound-sharing sources (2 fields): For open-class fields lacking clean opposites (*object* and *action*), pairs contrast distinct sources that share an acoustic behavior, such as "bird" and "cricket" for chirping.
3. Relational permutations (3 fields): Pairs invert the temporal order (*sequence*), simultaneity (*concurrency*), or prominence (*foreground-background*) of two identical events (e.g., "An engine idles, then revs" vs. "An engine revs, then idles").

Candidate values start from the validation study's contrast pairs, on which the judge was already tested, and are extended with data from several sources: frequent values from the corpus frequency table, wording common in other caption corpora but rare in AudioCaps, and language-model proposals used primarily for style descriptors (which are rare in the caption datasets). All candidate

pairs were manually vetted under two constraints: the contrast must be audibly distinguishable, and the values must fit the same template slot so swapping them alters only one phrase. Each value pair is rendered with several fixed object–action *anchors* mined from AudioCaps by extracting single-source events from captions (Appendix E). Anchors are selected to remain orthogonal to the property under test: for example, "an engine running" leaves volume open and anchors *loudness*, while "a metallic clank" predetermines *material* and cannot anchor it. The vocabulary covers everyday sound sources and non-speech vocalizations, excluding music and speech.

Prompts are instantiated by fixed templates that ensure minimal-pair identity up to the target value phrase (e.g., "An engine runs loudly" vs. "An engine runs faintly"). Templates emulate natural captions to avoid format artifacts (Ronchini et al., 2025). The resulting suite includes 96 value pairs across 126 anchors, 69 source/action groups, and 150 scenario pairs, yielding 100 prompts per field (2,500 total). We evaluate eleven TTA models spanning latent diffusion, flow matching, autoregressive, and masked-generation designs: TangoFlux (Hung et al., 2026), Tango 2 (Majumder et al., 2024), EzAudio (Hai et al., 2025), AudioX (Tian et al., 2025), AudioGen (Kreuk et al., 2023), MAGNeT (Ziv et al., 2024), AudioLDM 2 (Liu et al., 2024), Stable Audio Open (Evans et al., 2024), MMAudio (Cheng et al., 2025), and Stable Audio 3 (SFX and Medium) (Evans et al., 2026). Each model generates one clip per prompt.

5.2 Scoring

Each generated clip is tested with two probes: the validated field probe from Section 4.2 and the holistic *Match* baseline (adaptation rules in Appendix D). Binary questions are scored by the normalized probability of the target token (Kuan et al., 2026a), and multiple-choice questions by the probability mass on the target option (Robinson & Wingate, 2023). For *Contrast* and *Dual-pair*, each minimal pair shares a probe whose target answer shifts between clips; we compute a joint pair score as the product of the two clip probabilities, following Winoground's group formulation (Thrush et al., 2022). To establish a baseline for text-only bias, each question is additionally scored against Gaussian noise, yielding a fixed no-listening floor F shared across all models. Reported scores average over three seeds and permuted option orders (Appendix G).

The noise control establishes each field's floor F, but a field's achievable score is also limited by the judge. To estimate an empirical ceiling, we had five human raters evaluate 340 clips generated by the three top-performing models across all fields (protocol in Appendix J). For each field, the ceiling C is the judge's mean score on the human-confirmed clips, taken at the upper end of a bootstrap 95% interval over units and capped at 1. Most fields yield ceilings near 1, indicating that low model scores stem from TTA generation failures rather than judge failures. However, on *distance* and *location*, the judge scored human-confirmed clips at the level of the noise control (Table J.1), even though it detects both on real audio. We therefore exclude these two fields as judge-limited and evaluate models on the remaining 23 fields. Table 3 lists for each field the selected probe's leakage-scaled gain as its judge reliability, the floor F, the same probe's score on real recordings that contain the field, and the ceiling C; the *Match* row is a raw probability with its own floor.

Because floors and ceilings differ across fields, raw scores are not comparable between fields. Table 3 therefore reports each model's normalized gain $(S - F)/(C - F)$, which maps a score at the no-listening floor F to 0 and a score at the empirical ceiling C to 1; its reference columns give the raw floor, real-audio score, and ceiling that define this mapping. Our global aggregate averages normalized gain across the 23 valid fields (Table 3, bottom row), and the dimension aggregates of Figure 2 average over their constituent fields. Appendix G lists the raw scores of all 25 fields and details aggregate variants without the ceiling and reliability-weighted variants.

5.3 GASA Benchmark Results

On *Match*, the non-decomposed baseline, EzAudio and TangoFlux lead at .840 and .811 against a control of .048, and the same two models sit at or near the top of most field rows, the decomposed metrics. A close middle group runs from Tango 2 at .777 down to AudioLDM 2 at .680. Under the normalized gain, EzAudio at .388 and TangoFlux at .384 lead. Bootstrap confidence intervals over prompts (Appendix G) separate this ordering at most points but not at the top: the EzAudio-TangoFlux gap is +.004 with a 95% interval of [-.016, +.023], so the two leaders are statistically indistinguishable. Figure 2 shows the same scores aggregated to the seven dimensions.

Table 3: Normalized gain $(S - F)/(C - F)$ of the eleven text-to-audio models on the 23 scored fields; three best models per row shaded. The reference columns are raw scores defined in §5.2, and the *Match* row is a raw probability with its own control; raw scores of all 25 fields are in Table G.4. SA3, Stable Audio 3; SAO, Stable Audio Open.

Field	Probe	Reference				Models											
		Judge reliability	Control	Real audio	Ceiling	EzAudio	TangoFlux	Tango 2	AudioX	SA3 SFX	SA3 Medium	SAO	AudioGen	MMAudio	AudioLDM 2	MAGNeT	Mean
object	*Dual-pair*	.583	.005	.566	.974	.613	.618	.551	.589	.530	.508	.577	.519	**.689**	.385	.401	.544
action	*Dual-pair*	.498	.012	.459	.987	.459	.465	.365	.385	.463	.415	.403	.352	**.468**	.354	.211	.395
material	*Dual-pair*	.589	.008	.534	.655	.075	.192	.103	.148	.098	.149	.186	.176	**.270**	.094	.096	.144
count	*Complete*	.786	.557	.811	.999	**.502**	.452	.410	.467	.440	.435	.394	.400	.401	.442	.467	.437
human specifics	*Complete*	.893	.477	.906	.999	.571	.610	.602	.575	.542	.613	**.661**	.499	.630	.577	.516	.581
pitch	*Dual-pair*	.621	.033	.587	.943	.215	.066	.069	.161	.137	.216	**.239**	.058	.066	.143	.013	.126
loudness	*Dual-pair*	.655	.010	.610	.792	.146	.123	.057	**.152**	.119	.074	.096	.085	.044	.070	.088	.096
timbre	*Dual-pair*	.669	.012	.605	.618	**.214**	.155	.088	.211	.141	.101	.118	.123	.089	.072	.050	.124
duration	*Dual-pair*	.771	.015	.718	.832	**.189**	.091	.047	.127	.061	.049	.036	.049	.004	.156	.023	.076
repetition	*Complete*	.819	.598	.863	.999	**.563**	.475	.482	.526	.529	.513	.438	.448	.482	.463	.267	.471
direction	*Target*	.492	.237	.685	.768	.340	.320	.371	.360	.359	**.448**	.428	.209	.239	.270	.219	.324
loudness dynamics	*Contrast*	.739	.130	.851	.866	**.314**	.210	.064	.059	.048	.050	.056	.040	.061	.086	.052	.094
pitch dynamics	*Target*	.866	.011	.817	.936	.365	.463	.434	.404	.598	**.655**	.410	.360	.273	.252	.459	.425
timbre dynamics	*Target*	.727	.213	.726	.845	.529	.476	**.534**	.436	.409	.408	.354	.441	.256	.317	.493	.423
distance dynamics	*Target*	.907	.110	.907	.973	.669	.682	.727	.560	**.764**	.755	.638	.712	.402	.496	.683	.644
direction dynamics	*Target*	.921	.213	.895	.989	.317	.386	**.389**	.244	.318	.230	.223	.317	.237	.199	.350	.292
repetition dynamics	*Contrast*	.555	.135	.854	.452	**.302**	.125	.021	.219	.155	.211	.090	.050	.263	.005	-.191	.114
acoustic environment	*Contrast*	.863	.157	.906	.797	.266	.237	.230	.136	.152	.226	**.278**	.121	.276	.048	.084	.187
emotion	*Target*	.862	.013	.888	.975	.564	.560	.561	.585	.583	**.587**	.465	.453	.536	.552	.464	.537
style	*Target*	.652	.383	.776	.983	.495	.439	.407	.392	.499	.526	.534	.429	.551	**.648**	.555	.498
temporal sequence	*Target*	.645	.030	.787	.999	.734	**.965**	.760	.437	.312	.270	.291	.592	.290	.275	.477	.491
temporal concurrency	*Contrast*	.716	.123	.776	.873	.268	**.445**	.353	.203	.045	.069	.067	.145	.047	.038	.136	.165
foreground / background	*Contrast*	.637	.154	.798	.645	.217	**.276**	.107	.039	.020	.161	.232	.146	.209	.241	.003	.150
Mean normalized gain (23 fields)				–		**.388**	.384	.336	.322	.318	.333	.314	.292	.295	.269	.257	
Match (raw probability, all prompts)		.048		–	–	**.840**	.811	.777	.773	.742	.713	.729	.738	.702	.680	.724	

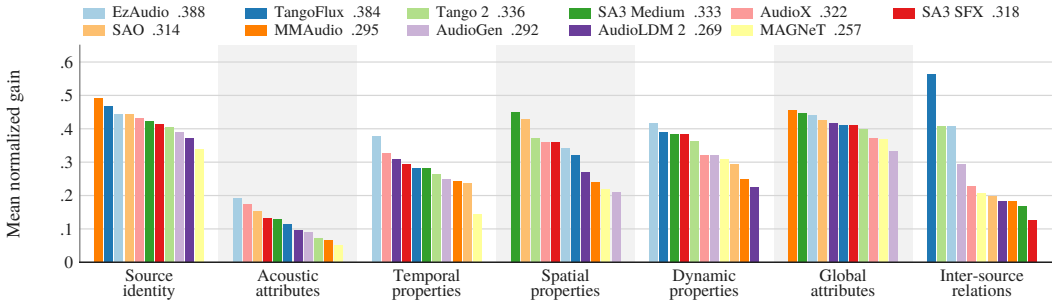

Figure 2: Normalized gain averaged over the fields of each dimension (distance and location left out); the number after each model is its 23-field average from Table 3.

Spread across fields is much larger than spread across models. At the strong end, temporal sequence, distance dynamics, object, human specifics, pitch dynamics, style, and emotion reach .59 to .96 normalized gain for their best models, peaking with TangoFlux on temporal sequence (.965). At the weak end, duration, loudness, loudness dynamics, repetition dynamics, pitch, and timbre reach on average only a tenth of their detectable range, indicating that current models do not reliably express these properties. Averaged over the models, source identity and the global attributes are the two strongest dimensions and static acoustic attributes by far the weakest; pitch and timbre are expressed far more reliably as changes than as static attributes. Among the relations, temporal sequence stands out while concurrency and foreground-background remain low.

6 DISCUSSION

No model leads everywhere, and newer is not always better. The two strongest models, EzAudio and TangoFlux, are statistically indistinguishable on the primary aggregate, and neither leads consistently across fields (EzAudio is ahead on thirteen of the 23 fields, TangoFlux on ten). Newer models are not consistently better. Stable Audio 3 improves little on Stable Audio Open in the aggregate (.318 and .333 against .314) but has a different profile, stronger on most dynamic attributes and weaker on static pitch and human specifics, so a similar total can hide different strengths; TangoFlux, although ahead of Tango 2 in the aggregate, trails it on seven fields.

Models render sounding objects better than actions. The best object scores exceed .6, while no model reaches .5 on action. One likely cause is that sources and typical behaviors frequently co-occur in captions, so a model may learn them as a unit; for example, 69% of AudioCaps captions mentioning a dog say it barks. The ontology also files some actions under a generic source, idling and revving under Engine rather than Car (Gemmeke et al., 2017), so a clip can carry the action label without one for the vehicle, and fewer examples tie the two together. Both effects make the object contrast easy and the action contrast hard: swapping bird for cricket also swaps their usual sounds, whereas swapping bark for whimper asks the dog for a sound that its captions rarely give it.

Models realize event order far better than concurrency or prominence. Audio-language models are reported to reason poorly over the relations between sound sources (Ghosh et al., 2024; Kuan & Lee, 2025). Our validation shows the same weakness for a whole-caption question, but a probe aimed at the relation recovers it (§4.2), and the human evaluation confirms that the judge detects all three relations on generated clips, order most sharply and foreground-background least (Table J.1). The generation results are split. Temporal sequence carries the highest score in the benchmark, while temporal concurrency and foreground-background are low for every model. Training recipes explain part of the split: the two models closest to TangoFlux on temporal sequence explicitly supervise event order, Tango 2 by optimizing preferences over order-rewritten captions (Majumder et al., 2024) and EzAudio by training on timestamped captions (Hai et al., 2025), whereas the models that score lowest on this field do not train on order. TangoFlux's own lead (.965 against EzAudio's .734) is surprising, because nothing in its training targets event order: its preference optimization ranks candidate clips by CLAP similarity to the prompt (Hung et al., 2026), a signal that is largely invariant to event order (Ghosh et al., 2024). A plausible source is its final training stages, which fine-tune on AudioCaps and draw the preference-optimization prompts from AudioCaps captions (Hung et al., 2026); in that corpus, temporal sequence is the most frequently marked field after object and action (27.7%, Table B.1). Among the models that score lowest on this field, the Stable Audio models build their prompts from file metadata (Evans et al., 2024; 2026), MMAudio learns timing from video rather than from text (Cheng et al., 2025), and AudioLDM 2 notes that CLAP conditioning loses temporal information and adds a FLAN-T5 encoder to compensate (Liu et al., 2024). Concurrency has received less direct supervision: AudioGen mixes pairs of training clips at random offsets (Kreuk et al., 2023), yet its concurrency score does not stand out, and Stable Audio Open's authors report that their model sometimes omits one of the sounds in prompts that join events with connectors (Evans et al., 2024). We suspect the difference lies in what each relation asks of the generator: an order can be realized by producing two events one after the other, whereas concurrency and foreground-background require two events to be audible at once at a controlled balance, which sequential placement does not provide.

Models render pitch and timbre better dynamically than statically and render loudness poorly for both. A likely reason is that a static value needs an external reference and a change does not. Repetition, a static field that scores far above its dynamic counterpart, needs no reference either. Loudness is the exception: its static field stays barely above the control for every model, as does its dynamic field for all but EzAudio and TangoFlux. Training data explains part of this: loudness is marked in only 12.7% of AudioCaps captions (and only with relative terms) while pitch and timbre are marked in under 5.0% (Appendix B). None of the models takes loudness or pitch as an input; the only acoustic quantities any take as input are duration and timing (Hung et al., 2026; Evans et al., 2024). On the output side, previous work found generated audio varies less in loudness and timbre than real recordings (Morse et al., 2025).

Takeaways for model development. Current models render what captions describe most often: which sources are present, what they usually do, which of two events comes first, and how pitch, timbre, or distance changes over the clip. They do not render what captions rarely specify: how loud, how high, or how long a sound is, what its timbre is, two sounds at once with one in front, and a source doing something it rarely does. Every architecture shares these gaps, so we read them as gaps in the training signal rather than in the models. Event order, the one relation trained for directly, through order-swapped preference pairs (Majumder et al., 2024) and captions written from timestamps (Hai et al., 2025), is the relation on which the leaders score highest; concurrency and prominence could be trained the same way if mixtures were labeled for them. Absolute loudness and pitch become controllable when given as conditions or stated in the caption (Wang et al., 2025; Flores García et al., 2025), but no model in our set receives them. The probes can also serve for training: Kuan et al. (2026b) build preference pairs from an audio-language model's judgments of event presence and order and improve TangoFlux's instruction following, and the probes validated here supply the same judgment for 23 fields.

7 LIMITATIONS

All of the validation data are drawn from AudioCaps, whose field distribution is highly uneven, and because there are not established, publicly released datasets of real user prompts, our 25-field hierarchy draws on several caption corpora together with established audio taxonomies and perception research. However, captions are not the same as prompts, and thus our dataset may not cover the full range of properties prompts specify. The probes also tie every attribute to the main source of a prompt. A prompt describes one source, or two sources in a stated relation, and each attribute question is about that main source, so the benchmark does not test whether a model gives the right attribute to the right source when two sources have different attributes. Future work could identify missing dimensions from real prompts and bring both them and cross-source binding under the same field-level treatment.

Our primary evaluation uses Qwen3-Omni, although rescoring with two other judges yields largely unchanged model rankings (Appendix H). The human evaluation of §5.2 anchors the judge's ceiling on generated clips, but it uses five raters and ten prompts per field from the three top-ranked models, so a ceiling is a point estimate with a wide interval on the fields that few raters confirmed, leaving unverified how the judge handles synthesis artifacts from weaker models. It also shows that on two fields, distance and location, the judge cannot detect the property on generated clips although it does on real recordings. Their raw scores (Table G.4) measure the judge rather than the generators, and a judge that hears spatial properties in synthetic audio is needed before the benchmark can score them. Finally, the structured representations depend on the language model that annotated them; while constrained by explicit-evidence extraction (Appendix B), reliability remains bounded by the underlying model.

8 CONCLUSION

This paper recasts text-to-audio alignment from a single global property into a structured one that holds or fails separately for each field a prompt specifies. We organized the properties a scene description can express into 25 fields, built probes for each field from real caption-audio pairs, validated them under real audio and under a Gaussian-noise control, and selected Qwen3-Omni from a comparison of three audio-language judges. The resulting judge and probe suite underpin the Generative Audio Structured Alignment Benchmark, which evaluates eleven current text-to-audio models on controlled minimal-pair prompts and shows how far structured alignment remains from solved: even the leaders realize only part of what their prompts specify, and scores differ far more between fields than between models. The per-field profile it reports for each model shows where the generated audio follows the prompt and where it departs from it, a distinction that a single alignment score cannot draw.

9 AI USE STATEMENT

We used generative AI tools to generate synthetic data (Qwen2.5-72B-Instruct annotates captions and prompts against the field hierarchy, and a language model proposed candidate prompt values, which we checked by hand). We did not use generative AI tools to develop conceptual frameworks, propose or refine hypotheses, or interpret results, and formulating or proving mathematical claims, translation, and qualitative analysis are not applicable to this work. Additionally, we used generative AI tools to edit code, create scientific figures, edit the paper for grammar and readability, and review it for unclear statements. We have reviewed all AI-assisted work: AI-assisted code was tested by the authors, and all text was revised by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

10 REPRODUCIBILITY STATEMENT

All eleven text-to-audio models, the three LALM judges (Qwen3-Omni, Qwen2.5-Omni, AudioFlamingo-Next), and the annotation model (Qwen2.5-72B-Instruct) are open models used with their released weights. The caption-audio pairs of the validation study and the anchors of the benchmark prompts come from AudioCaps, and the human-rating comparison uses the public RELATE dataset. The field hierarchy, annotation rules, probe formats and judge prompts, value and anchor tables, aggregation protocol, and human evaluation protocol are given in Appendices A–J. The benchmark prompts, the judge questions, the generation, scoring, and aggregation code, and the result tables of all eleven models are available in an anonymized repository at `https://anonymous.4open.science/r/gasa-benchmark-963B/`; the per-prompt judge scores will be added to the same repository.

REFERENCES

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. MMAudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Yoonjin Chung, Pilsun Eu, Junwon Lee, Keunwoo Choi, Juhan Nam, and Ben Sangbae Chon. KAD: No more FAD! an effective and efficient evaluation metric for audio generation. *arXiv preprint arXiv:2502.15602*, 2025.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740, 2020. doi: 10.1109/ICASSP40776.2020.9052990.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095889.

Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. Natural language supervision for general-purpose audio representations. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. *arXiv preprint arXiv:2407.14358*, 2024.

Zach Evans, Julian D. Parker, Matthew Rice, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio 3. *arXiv preprint arXiv:2605.17991*, 2026.

Hugo Flores García, Oriol Nieto, Justin Salamon, Bryan Pardo, and Prem Seetharaman. Sketch2Sound: Controllable audio generation via time-varying signals and sonic imitations. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888184.

William W. Gaver. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological Psychology*, 5(1):1–29, 1993. doi: 10.1207/s15326969eco0501_1.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017. doi: 10.1109/ICASSP.2017.7952261. URL https://research.google/pubs/pub45857.

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Reddy Evuru, Ramaneswaran S, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. CompA: Addressing the gap in compositional reasoning in audio-language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

Sreyan Ghosh, Arushi Goel, Kaousheik Jayakumar, Lasha Koroshinadze, Nishit Anand, Zhifeng Kong, et al. Audio Flamingo Next: Next-generation open audio-language models for speech, sound, and music. *arXiv preprint arXiv:2604.10905*, 2026.

Catherine Guastavino. Everyday sound categorization. In *Computational Analysis of Sound Scenes and Events*, pp. 183–213. Springer International Publishing, Cham, 2018. doi: 10.1007/978-3-319-63450-0_7.

Tianyu Guo, Hongyu Chen, Hao Liang, Meiyi Qiang, Bohan Zeng, Linzhuang Sun, Bin Cui, and Wentao Zhang. BRACE: A benchmark for robust audio caption quality evaluation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2512.10403.

Jiarui Hai, Yong Xu, Hao Zhang, Chenxing Li, Helin Wang, Mounya Elhilali, and Dong Yu. EzAudio: Enhancing text-to-audio generation with efficient diffusion transformer. In *Proceedings of Interspeech*, pp. 4233–4237, 2025.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A. Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, 2023.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Ali Bagherzadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. TangoFlux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

Yusuke Kanamori, Yuki Okamoto, Taisei Takano, Shinnosuke Takamichi, Yuki Saito, and Hiroshi Saruwatari. RELATE: Subjective evaluation dataset for automatic evaluation of relevance between text and audio. In *Proceedings of Interspeech*, pp. 3155–3159, 2025. URL https://arxiv.org/abs/2506.23582.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proceedings of Interspeech*, pp. 2350–2354, 2019.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 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*, pp. 119–132, 2019. doi: 10.18653/v1/N19-1011. URL https://aclanthology.org/N19-1011/.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. AudioGen: Textually guided audio generation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Chun-Yi Kuan and Hung-yi Lee. Can large audio-language models truly hear? Tackling hallucinations with multi-task assessment and stepwise audio reasoning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

Chun-Yi Kuan, Kai-Wei Chang, and Hung-yi Lee. AQAScore: Evaluating semantic alignment in text-to-audio generation via audio question answering. *arXiv preprint arXiv:2601.14728*, 2026a.

Chun-Yi Kuan, Siwon Kim, Byeonggeun Kim, Suyoun Kim, Bo-Ru Lu, Qingming Tang, Ankur Gandhe, Hung yi Lee, Chieh-Chi Kao, and Chao Wang. Improving text-to-audio instruction following via fine-grained feedback from audio-aware large language models. In *Proceedings of Interspeech*, 2026b.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th ACM Symposium on Operating Systems Principles*, pp. 611–626, 2023. doi: 10.1145/3600006.3613165. URL https://arxiv.org/abs/2309.06180.

Tiep Le, Vasudev Lal, and Phillip Howard. COCO-Counterfactuals: Automatically constructed counterfactual examples for image-text pairs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 71195–71221, 2023.

Huan Liao, Haonan Han, Kai Yang, Tianjiao Du, Rui Yang, Qinmei Xu, Zunnan Xu, Jingquan Liu, Jiasheng Lu, and Xiu Li. BATON: Aligning text-to-audio model using human preference feedback. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4542–4550, 2024.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision (ECCV)*, 2024.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21450–21474. PMLR, 2023.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. AudioLDM 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024. doi: 10.1109/TASLP.2024.3399607.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM)*, pp. 564–572, 2024. doi: 10.1145/3664647.3681688.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. WavCaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024. doi: 10.1109/TASLP.2024.3419446. URL https://arxiv.org/abs/2303.17395.

Jonathan Morse, Azadeh Naderi, Swen Gaudl, Mark Cartwright, Amy K. Hoover, and Mark J. Nelson. Expressive range characterization of open text-to-audio models. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 21(1):91–98, 2025. doi: 10.1609/aiide.v21i1.36813.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Joshua Robinson and David Wingate. Leveraging large language models for multiple choice question answering. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

Francesca Ronchini, Ho-Hsiang Wu, Wei-Cheng Lin, and Fabio Antonacci. Mind the prompt: Prompting strategies in audio generations for improving sound classification. In *Proceedings of the Generative Data Augmentation for Real-World Signal Processing Applications (GenDA) Workshop at ICASSP 2025*, 2025.

Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2V-CompBench: A comprehensive benchmark for compositional text-to-video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8406–8416, 2025.

Shuntaro Suzuki, Kento Tokura, Daichi Yashima, Kanon Amemiya, Komei Sugiura, and Shinnosuke Takamichi. ELSA: Acoustic event-level semantic alignment for fine-grained reference-free text-to-audio evaluation. *arXiv preprint arXiv:2606.17404*, 2026.

Taisei Takano, Yuki Okamoto, Yusuke Kanamori, Yuki Saito, Ryotaro Nagase, and Hiroshi Saruwatari. Human-CLAP: Human-perception-based contrastive language-audio pretraining. In *Proceedings of the Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 131–136, 2025. doi: 10.1109/APSIPAASC65261.2025.11249121.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Zeyue Tian, Yizhu Jin, Zhaoyang Liu, Ruibin Yuan, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. AudioX: Diffusion transformer for anything-to-audio generation. *arXiv preprint arXiv:2503.10522v1*, 2025.

Hui Wang, Cheng Liu, Junyang Chen, Haoze Liu, Yuhang Jia, Shiwan Zhao, Jiaming Zhou, Haoqin Sun, Hui Bu, and Yong Qin. TTA-Bench: A comprehensive benchmark for evaluating text-to-audio models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33512–33520, 2026. doi: 10.1609/aaai.v40i39.40639. URL https://arxiv.org/abs/2509.02398.

Yuanyuan Wang, Hangting Chen, Dongchao Yang, Zhiyong Wu, and Xixin Wu. AudioComposer: Towards fine-grained audio generation with natural language descriptions. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888303.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095969.

Tianxin Xie, Wentao Lei, Kai Jiang, Guanjie Huang, Pengfei Zhang, Chunhui Zhang, Fengji Ma, Haoyu He, Han Zhang, Jiangshan He, Jinting Wang, Linghan Fang, Lufei Gao, Orkesh Ablet, Peihua Zhang, Ruolin Hu, Shengyu Li, Weilin Lin, Xiaoyang Feng, Xinyue Yang, Yan Rong, Yanyun Wang, Zihang Shao, Zelin Zhao, Chenxing Li, Shan Yang, Wenfu Wang, Meng Yu, Dong Yu, and Li Liu. PhyAVBench: A challenging audio physics-sensitivity benchmark for physically grounded text-to-audio-video generation. *arXiv preprint arXiv:2512.23994*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, et al. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Yongyi Zang, Sean O'Brien, Taylor Berg-Kirkpatrick, Julian J. McAuley, and Zachary Novack. Are you really listening? boosting perceptual awareness in music-QA benchmarks. In *Proceedings of the 26th International Society for Music Information Retrieval Conference (ISMIR)*, 2025. URL https://arxiv.org/abs/2504.00369.

Alon Ziv, Itai Gat, Gael Le Lan, Tal Remez, Felix Kreuk, Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. Masked audio generation using a single non-autoregressive transformer. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.

APPENDIX

A FULL FIELD HIERARCHY

This appendix gives the operational definition of each of the 25 fields. A field is marked present only when the text contains explicit wording for it. Table A.1 gives the definition and an example caption for each field, with the wording that expresses the field in bold.

Table A.1: Definition and example caption for each field. The wording that expresses the field is in bold.

Dimension	Field	Definition	Example
Source identity	object	The physical entity producing the sound. Marked when the text names an identifiable source, such as "dog" in "a dog barking," not for generic wording such as "a sound is heard."	A **woman** is speaking while dishes clatter.
	action	The physical action producing the sound. Marked when the text contains a verb or event noun for sound-producing activity, such as barking, knocking, or pouring.	Someone **knocks** on a wooden door.
	material	The material of the sounding body. Marked only when a material or surface type is named, such as wooden, metal, or glass.	Rain hits a **metal** roof during a storm.
	count	The number or quantity of sound-producing entities. Marked for numbers and quantity words such as two, several, or flock, not for a bare article such as "a."	**Two** dogs bark near a fence.
	human specifics	Demographic or physiological attributes of a human source, such as woman, child, or elderly man. Generic person references are not marked.	A **child** laughs while adults talk nearby.
Acoustic attributes	pitch	Explicit frequency-height descriptors such as high-pitched, deep, or low.	A **high-pitched** siren approaches from the distance.
	loudness	Explicit static volume descriptors such as loud, faint, or soft. Changes such as getting louder are marked under loudness dynamics.	A **faint** voice speaks under traffic noise.
	timbre	Spectral or tonal character beyond pitch and loudness, such as harsh, muffled, bright, or raspy.	A **muffled** voice comes from behind a closed door.
Temporal properties	duration	Explicit length or persistence of a sound event, such as brief, sustained, or long.	A **brief** horn sounds before the engine starts.
	repetition	Explicit repetitive pattern, rhythm, or pacing, such as rhythmic, intermittent, repeated, or fast tapping.	**Repeated** tapping continues beside running water.
Spatial properties	distance	Explicit proximity between source and listener, such as distant, nearby, or far away.	**Distant** thunder rolls as rain falls.
	direction	Explicit directional position relative to the listener, such as left, overhead, or behind.	A voice comes from **below** as music plays.

Dimension	Field	Definition	Example
Dynamic properties	loudness dynamics	Volume changing over time, such as gradually getting louder, fading out, or dying down.	A siren approaches, **becoming louder**.
	pitch dynamics	Pitch changing over time, such as a rising pitch or a descending whistle.	A whistle shifts into a **rising pitch**.
	timbre dynamics	Timbral quality changing over time, such as a sound becoming more muffled.	The voice **becomes more muffled** as the door closes.
	distance dynamics	Distance changing over time, such as approaching, receding, or fading into the distance.	A car **approaches** with its engine growing clearer.
	direction dynamics	Direction changing over time, such as moving from left to right.	A bird flies **from left to right** overhead.
	repetition dynamics	Repetition rate or pattern changing over time, such as accelerating footsteps or increasingly irregular knocking.	Footsteps **gradually get faster** in the hallway.
Global attributes	acoustic environment	The acoustic space that shapes reverberation or echo, such as a cathedral, bathroom, open field, or outdoors.	Voices **echo in a large hall**.
	location	The real-world setting or place of the scene, such as a busy street, forest, or construction site.	Cars honk on a **busy street**.
	emotion	An explicit emotional or mood descriptor of the scene, such as eerie, cheerful, tense, or calm.	An **eerie** atmosphere surrounds the distant bells.
	style	An explicit aesthetic style, genre, or production descriptor, such as cinematic, 8-bit, or retro.	An **8-bit** melody plays over arcade sounds.
Inter-source relations	temporal sequence	Explicit ordering of events, such as one sound followed by another or one event occurring after another.	A dog barks, **then** a bell rings.
	temporal concurrency	Explicit simultaneity, such as one sound occurring while another happens or one sound accompanying another.	A woman talks **while** water pours.
	foreground-background	Explicit prominence relation between sources, such as one sound in the foreground and another in the background.	A man speaks with traffic in the **background**.

B Annotation and Coverage

Annotation Rules

The 25-field annotation is performed by Qwen2.5-72B-Instruct (Qwen Team, 2024), served with vLLM (Kwon et al., 2023) and guided JSON decoding. Each response contains a short reasoning string followed by one boolean per field. The reasoning string is required to quote the wording that triggered each marked field, which makes the annotation auditable.

The annotation rules enforce four constraints:

1. A field is marked only when the text contains words that directly express it.
2. World knowledge must not fill in unstated fields.
3. Logical implication is not treated as field presence.
4. Every marked field must be backed by quoted evidence.

Five fixed validation captions serve as a self-test of these rules:

- "A woman talks nearby as water pours" marks object, action, distance, and temporal concurrency, but not loudness, timbre, or acoustic environment.

- "A loud explosion followed by debris falling on metal" marks object, action, loudness, material, and temporal sequence, but not loudness dynamics from "explosion" alone.
- "A dog barking" marks object and action only.
- "Rhythmic ticking of a clock in a quiet library, gradually getting faster" marks object, action, repetition, loudness, location, and repetition dynamics.
- "A high-pitched siren approaching from the distance, becoming louder" marks object, pitch, distance, distance dynamics, and loudness dynamics, but not pitch dynamics.

Bottom-Up Taxonomy Development

Before the 25-field hierarchy was fixed, a bottom-up pass derived an 18-field list from caption corpora. The list grouped properties into objects and actions, acoustic features, spatial properties, temporal properties, and scene-level attributes. That pass already showed the distributional pattern motivating the final hierarchy: source identity and action are frequent, while many acoustic, dynamic, and relational properties are rare. The 25-field hierarchy consolidated this evidence with audio taxonomies and compositional benchmark factors, splitting dynamic change into per-attribute fields, separating inter-source relations, and tightening all definitions to the explicit-evidence standard.

Per-Dataset Field Coverage

The following table reports the number and percentage of captions in which each field is marked present under the 25-field rules, across AudioCaps (Kim et al., 2019), Clotho (Drossos et al., 2020), WavCaps (Mei et al., 2024), RELATE (Kanamori et al., 2025), and BRACE (Guo et al., 2025).

Table B.1: Per-dataset field coverage. Each cell gives the number of captions in which the field is marked present under the 25-field annotation rules, with the within-corpus percentage in parentheses.

Field	**AudioCaps** 57,188	**Clotho** 29,645	**WavCaps** 16,196	**RELATE** 1,927	**BRACE** 2,841	**Total** 107,797
object	53,871 (94.2)	28,333 (95.6)	14,970 (92.4)	1,847 (95.8)	2,711 (95.4)	101,732 (94.4)
action	54,921 (96.0)	28,522 (96.2)	13,698 (84.6)	1,868 (96.9)	2,733 (96.2)	101,742 (94.4)
temporal sequence	15,819 (27.7)	4,937 (16.7)	641 (4.0)	991 (51.4)	727 (25.6)	23,115 (21.4)
temporal concurrency	11,690 (20.4)	6,237 (21.0)	2,222 (13.7)	375 (19.5)	686 (24.1)	21,210 (19.7)
human specifics	13,037 (22.8)	2,097 (7.1)	1,644 (10.2)	458 (23.8)	396 (13.9)	17,632 (16.4)
loudness	7,255 (12.7)	3,620 (12.2)	243 (1.5)	231 (12.0)	302 (10.6)	11,651 (10.8)
foreground-background	5,431 (9.5)	3,833 (12.9)	1,804 (11.1)	222 (11.5)	319 (11.2)	11,609 (10.8)
repetition	4,880 (8.5)	5,511 (18.6)	626 (3.9)	146 (7.6)	373 (13.1)	11,536 (10.7)
count	4,866 (8.5)	3,124 (10.5)	630 (3.9)	191 (9.9)	296 (10.4)	9,107 (8.4)
material	2,813 (4.9)	4,764 (16.1)	750 (4.6)	177 (9.2)	344 (12.1)	8,848 (8.2)
location	964 (1.7)	4,287 (14.5)	1,747 (10.8)	20 (1.0)	285 (10.0)	7,303 (6.8)
distance	4,478 (7.8)	1,763 (5.9)	597 (3.7)	147 (7.6)	205 (7.2)	7,190 (6.7)
timbre	2,825 (4.9)	2,178 (7.3)	574 (3.5)	96 (5.0)	174 (6.1)	5,847 (5.4)
loudness dynamics	1,352 (2.4)	1,689 (5.7)	66 (0.4)	60 (3.1)	134 (4.7)	3,301 (3.1)
distance dynamics	1,486 (2.6)	1,189 (4.0)	198 (1.2)	50 (2.6)	119 (4.2)	3,042 (2.8)
acoustic environment	607 (1.1)	1,411 (4.8)	530 (3.3)	24 (1.2)	98 (3.4)	2,670 (2.5)
duration	1,499 (2.6)	709 (2.4)	97 (0.6)	49 (2.5)	69 (2.4)	2,423 (2.2)
pitch	1,141 (2.0)	597 (2.0)	122 (0.8)	36 (1.9)	44 (1.5)	1,940 (1.8)
repetition dynamics	528 (0.9)	913 (3.1)	74 (0.5)	15 (0.8)	57 (2.0)	1,587 (1.5)
direction	419 (0.7)	832 (2.8)	173 (1.1)	9 (0.5)	29 (1.0)	1,462 (1.4)
emotion	346 (0.6)	271 (0.9)	214 (1.3)	14 (0.7)	18 (0.6)	863 (0.8)
direction dynamics	238 (0.4)	353 (1.2)	64 (0.4)	7 (0.4)	27 (1.0)	689 (0.6)
style	165 (0.3)	186 (0.6)	265 (1.6)	11 (0.6)	9 (0.3)	636 (0.6)
pitch dynamics	111 (0.2)	276 (0.9)	26 (0.2)	4 (0.2)	15 (0.5)	432 (0.4)
timbre dynamics	60 (0.1)	114 (0.4)	29 (0.2)	5 (0.3)	6 (0.2)	214 (0.2)

C Full Validation Matrix

This appendix reports the full validation matrix behind the judge comparison of Section 4.2, every field-probe combination for each of the three judges. Each cell gives the leakage-scaled gain defined in Section 4.2. Cell sample sizes are identical across the judges because the same probe questions were scored by all three, so they are shown once, in the Qwen3-Omni table. A dash marks a combination that cannot be constructed.

We also conducted a split-half check to address the issue that selecting each field's probe using the same validation data that scores it can overstate the probe's gain. Our check showed that selecting on half of the items and scoring on the held-out half lowered the mean selected gain by only .058. In addition, the same probe was selected in at least 75% of the splits across 16 of the 25 fields. The number of probe questions reflects each field's natural occurrence in AudioCaps, ranging from 26 for rare attributes like timbre dynamics to several hundred for common fields.

Table C.1: Leakage-scaled gain under Qwen2.5-Omni.

Field	Match	Contrast	Target	Complete	Dual-pair
object	.432	—	.419	.550	.464
action	.337	—	.210	.563	.344
material	.574	.064	.111	.215	.380
count	.673	.007	.158	.077	.608
human specifics	.423	.647	.222	.410	.720
pitch	.558	.208	.216	.154	.326
loudness	.710	.169	.252	.092	.440
timbre	.607	.091	.103	.118	.482
duration	.673	.079	.079	.103	.552
repetition	.732	.297	.473	.182	.596
distance	.647	.074	.078	.000	.479
direction	.421	.000	.148	.103	.222
loudness dynamics	.557	.213	.380	.278	.591
pitch dynamics	.583	.400	.362	.250	.256
timbre dynamics	.630	.333	.280	.579	.357
distance dynamics	.585	.244	.586	.164	.473
direction dynamics	.574	.189	.461	.091	.405
repetition dynamics	.593	.340	.370	.319	.375
acoustic environment	.531	.114	.276	.219	.617
location	.545	.153	.563	.296	.565
emotion	.464	.250	.314	.267	.293
style	.519	.240	.234	.293	.500
temporal sequence	.219	.024	.269	.379	.031
temporal concurrency	.108	.094	.067	.376	.031
foreground-background	.006	.000	.187	.376	—

Table C.2: Leakage-scaled gain under Qwen3-Omni, with per-cell sample sizes in parentheses.

Field	Match	Contrast	Target	Complete	Dual-pair
object	.509 (876)	—	.420 (876)	.546 (876)	.583 (386)
action	.406 (630)	—	.165 (630)	.410 (629)	.498 (289)
material	.704 (108)	.125 (2107)	.253 (1037)	.580 (260)	.589 (185)
count	.818 (110)	.620 (3455)	.556 (2143)	.786 (418)	.780 (346)
human specifics	.788 (104)	.476 (12109)	.789 (5526)	.893 (859)	.842 (792)
pitch	.660 (104)	.309 (777)	.540 (428)	.292 (130)	.621 (95)
loudness	.852 (120)	.293 (5966)	.432 (1728)	.629 (499)	.655 (522)
timbre	.814 (112)	.141 (1433)	.205 (800)	.309 (136)	.669 (142)
duration	.764 (110)	.591 (984)	.610 (761)	.641 (156)	.771 (166)
repetition	.857 (112)	.568 (2354)	.355 (874)	.819 (341)	.777 (256)
distance	.863 (102)	.002 (4196)	.066 (1026)	.478 (366)	.499 (353)
direction	.458 (114)	.000 (68)	.492 (116)	.179 (78)	.306 (36)

Field	Match	Contrast	Target	Complete	Dual-pair
loudness dynamics	.623 (154)	.739 (835)	.453 (395)	.649 (108)	.568 (88)
pitch dynamics	.688 (96)	.400 (20)	.866 (95)	.000 (50)	.385 (39)
timbre dynamics	.889 (54)	.667 (4)	.727 (26)	.158 (38)	.571 (14)
distance dynamics	.615 (130)	.526 (689)	.907 (1002)	.446 (110)	.484 (93)
direction dynamics	.556 (108)	.252 (127)	.921 (128)	.000 (53)	.541 (37)
repetition dynamics	.538 (104)	.555 (353)	.433 (428)	.418 (91)	.538 (65)
acoustic environment	.813 (128)	.863 (246)	.391 (624)	.356 (73)	.667 (60)
location	.782 (110)	.093 (236)	.534 (1248)	.526 (106)	.565 (69)
emotion	.911 (112)	.788 (208)	.862 (133)	.533 (98)	.512 (41)
style	.732 (108)	.400 (50)	.652 (77)	.452 (82)	.500 (30)
temporal sequence	.159 (4702)	.545 (2351)	.645 (4702)	.542 (4702)	.585 (159)
temporal concurrency	.019 (6248)	.716 (3120)	.546 (6248)	.697 (6244)	.374 (324)
foreground-background	.108 (1740)	.637 (870)	.520 (1740)	.583 (1740)	—

Table C.3: Leakage-scaled gain under AudioFlamingo-Next.

Field	Match	Contrast	Target	Complete	Dual-pair
object	.592	—	.502	.308	.397
action	.394	—	.179	.422	.244
material	.833	.187	.312	.108	.380
count	.818	.300	.781	.077	.581
human specifics	.885	.730	.802	.070	.659
pitch	.604	.201	.197	.015	.347
loudness	.867	.561	.152	.024	.411
timbre	.821	.000	.393	.000	.338
duration	.831	.303	.389	.000	.503
repetition	.732	.456	.391	.111	.644
distance	.824	.000	.107	.000	.318
direction	.500	.000	.213	.000	.194
loudness dynamics	.738	.424	.540	.148	.534
pitch dynamics	.646	.200	.594	.353	.256
timbre dynamics	.852	.000	.400	.105	.286
distance dynamics	.738	.548	.745	.000	.430
direction dynamics	.574	.488	.908	.189	.167
repetition dynamics	.603	.555	.504	.088	.385
acoustic environment	.781	.187	.554	.000	.400
location	.764	.102	.510	.075	.449
emotion	.768	.779	.794	.329	.195
style	.741	.600	.638	.238	.500
temporal sequence	.305	.507	.444	.088	.312
temporal concurrency	.040	.543	.567	.027	.130
foreground-background	.043	.195	.124	.048	—

D PROBE FORMATS

This appendix gives the question format posed to the judge for each probe family and works one instance through all five probes. For each marked field, which we call an instance, the annotation keeps two pieces of text. The canonical value is one label for all wordings of the same value: “far away,” “in the distance,” and “distant” all become “far”, so one test applies to every caption with that value. The evidence span is the original wording, such as “in the distance,” so that an edit to the caption changes only those words. An instance is kept only if its value is explicit in the text and makes an audible difference. A caption may say that a car passes on Main Street, but the street name cannot be heard, so it yields no instance. Table D.1 poses all five probes for the instance whose supported caption is “A dog barks in the distance”, with target field distance, evidence span “in the distance”, canonical value far, and contrast alternative nearby.

Table D.1: The five probes, as posed for one distance instance.

Probe	Probe question	Correct answer
Supported caption: *A dog barks in the distance*. Target field: distance.		
Match	"Does the following caption accurately describe the audio? *A dog barks in the distance*"	Yes
	"Does the following caption accurately describe the audio? *A train passes nearby*" (a caption from another clip)	No
Contrast	"Which caption more accurately describes the audio? A. *A dog barks in the distance* B. *A dog barks nearby*"	A
Target	"Does *a dog barking in the distance* accurately describe the distance aspect of this audio?"	Yes
	"Does *a dog barking nearby* accurately describe the distance aspect of this audio?"	No
Complete (remove)	"Which caption more accurately and completely describes the audio? A. *A dog barks* B. *A dog barks in the distance*"	B
Complete (add)	"Which caption more accurately and completely describes the audio? A. *A dog barks in the distance* B. *A dog barks in the distance while a car horn sounds*"	A
Dual-pair (clip 1)	"Which caption most accurately and completely describes the audio? A. *A dog barks in the distance* B. *A train passes nearby* C. *A dog barks in the distance while a car horn sounds* D. *A dog barks*"	A
Dual-pair (clip 2)	(the same four options in the same order, with only the audio changed)	B

Match. Poses the complete caption for yes/no verification.

```
Listen to the audio. Does the following caption accurately describe it?

"{caption}"

Please answer yes or no.
```

Target. Poses a yes/no question about one field value.

```
Listen to the audio. Does "{value phrase}" accurately describe the
  {field} aspect of this audio?

Please answer yes or no.
```

Contrast. Asks which caption more accurately describes the audio.

```
Listen to the audio. Which caption more accurately describes the audio?

A: {option 1}
B: {option 2}

Please answer with the letter only (A or B).
```

Complete. Asks which caption more accurately and completely describes the audio, in the same two-option format.

```
Listen to the audio. Which caption more accurately and completely
  describes the audio?

A: {option 1}
B: {option 2}

Please answer with the letter only (A or B).
```

Dual-pair. Uses the four-option format below, posed once for each clip of the pair.

```
Listen to the audio. Which caption most accurately and completely
  describes the audio?

A: {option}
B: {option}
C: {option}
D: {option}

Please answer with the letter only.
```

The distractor options pair the supported caption with a version that removes the target field, a version that adds an unsupported detail, and captions from another source clip. In the dual-audio variant, the same text options are posed with two clips; the correct option changes with the audio while the text remains fixed.

For the validation probes, the edited captions are produced in two ways. For the twenty fields with opposing values, two edits are made by rule on the evidence span: *Contrast* replaces the value with its opposite, and the remove variant of *Complete* deletes the span. Two edits are written by the language model that annotates the captions, Qwen2.5-72B-Instruct: the phrase that *Target* asks about, and the detail that the add variant of *Complete* inserts. The same model then checks every edited caption against fixed criteria and discards the ones that fail, and we check a sample of the remaining probes by hand. For object, action, and the three inter-source relations, no language model is involved: the phrases follow templates, and the added details are taken from a fixed list of eight.

Benchmark Adaptation

For scoring generated audio, *Match* and *Target* keep only their positive questions. *Complete* keeps its equal mix of add and remove variants, with added details drawn from the anchor distribution. *Contrast* offers the pair's two captions as options, and the *Dual-pair* adds two distractors, a pair caption with an added event and the caption with the target field removed.

E Benchmark Value and Anchor Tables

This appendix lists the reviewed tables from which the benchmark prompts of Section 5 are rendered, the value pairs for the twenty fields with opposing values, the object and action groups, the scenario pairs for the three inter-source relation fields, and the anchor pool. Every row was reviewed by hand as described in Section 5.1. For human specifics, the sounding person is itself the tested value, and its anchors are written ⟨person⟩.

Table E.1: Value pairs for the twenty fields with opposing values.

Field	**Value A**	**Value B**
material	metal	wood
material	ceramic	stone
material	glass	plastic
count	a few	lots of
count	three	a group of
count	two	many
count	a single	several
count	a pair of	a dozen
human specifics	adult	baby
human specifics	boy	girl
human specifics	elderly	young
human specifics	female	male
human specifics	man	woman
human specifics	adult	child
human specifics	lady	man

Field	Value A	Value B
human specifics	adult	infant
human specifics	adult	kid
pitch	deep	piercing
pitch	high frequency	low frequency
pitch	high-pitched	low-pitched
loudness	deafening	faint
loudness	loudly	quietly
loudness	noisily	softly
loudness	blaring	soft
loudness	booming	gentle
loudness	loud	quiet
timbre	clear	muffled
timbre	crisp	dull
timbre	mellow	harsh
timbre	smooth	raspy
timbre	buzzing	squeaky
duration	brief	prolonged
duration	briefly	for a long time
duration	long	short
duration	long-lasting	short-lived
duration	momentary	sustained
repetition	continuously	over and over
repetition	once	three times
repetition	twice	several times
repetition	once	repeatedly
repetition	constantly	intermittently
distance	close	distant
distance	close by	far away
distance	far off	up close
distance	in the distance	nearby
distance	a long distance away	a short distance away
direction	above	below
direction	behind	in front
direction	from above	from below
direction	from the left	from the right
loudness dynamics	fading in	fading out
loudness dynamics	gets louder	gets quieter
loudness dynamics	building up	dies down
pitch dynamics	ascending	descending
pitch dynamics	decreasing in pitch	increasing in pitch
pitch dynamics	falling in pitch	rising in pitch
pitch dynamics	gets higher	gets lower
pitch dynamics	getting higher	getting lower
timbre dynamics	becoming distorted	becoming crisper
timbre dynamics	becoming harsher	becoming smoother
timbre dynamics	brightening	darkening
timbre dynamics	becoming clearer	becoming muffled
distance dynamics	fading into the distance	drawing near
distance dynamics	getting closer	moving away
distance dynamics	approaching	going away
direction dynamics	moving around	moving back and forth
direction dynamics	passing by	staying in place
repetition dynamics	faster and faster	slower and slower
repetition dynamics	at irregular intervals	at regular intervals
repetition dynamics	getting faster	getting slower
repetition dynamics	without pause	starting and stopping
repetition dynamics	slowing down	speeding up
acoustic environment	in a parking garage	in a carpeted room
acoustic environment	in a tunnel	in an open field
acoustic environment	indoor	outdoor
acoustic environment	large hall	small room
acoustic environment	with a faint echo	with a strong echo
acoustic environment	enclosed	open space

Field	Value A	Value B
location	by a city street	by a highway
location	kitchen	farm
location	park	street
location	at a train station	in a restaurant
location	beach	forest
location	in an office	at a playground
location	in a church	at a construction site
emotion	angry	calm
emotion	anxious	cheerful
emotion	excited	fearful
emotion	happy	sad
emotion	joyful	mournful
emotion	playful	solemn
emotion	fearful	joyful
emotion	calm	frantic
style	horror	cartoon
style	sci-fi	realistic
style	cinematic	8-bit retro

Table E.2: Object and action groups. An object group lists sound sources that share the action that names it, and an action group lists actions attested for the source that names it.

Kind	Group	Members
object group	bang	firework, hammer
object group	beat	drum, heart
object group	beep	car, microwave, phone, truck
object group	boom	bass drum, thunder
object group	buzz	bee, doorbell, engine, fly, machine, motor
object group	chime	clock, doorbell
object group	chirp	bird, cricket, frog, insect, smoke detector, sparrow
object group	clatter	dish, typewriter
object group	click	heels, keyboard
object group	coo	baby, pigeon
object group	crack	ice, whip
object group	crash	cymbal, wave
object group	flap	bird, flag
object group	fly	aircraft, bee, helicopter, insect
object group	grunt	man, pig
object group	gurgle	baby, drain
object group	hiss	cat, snake, steam, tire
object group	honk	car, goose
object group	howl	wind, wolf
object group	hum	person, refrigerator, traffic, train
object group	play	cello, guitar, piano, violin
object group	ring	church bell, telephone
object group	roar	crowd, engine, jet, lion
object group	rumble	stomach, thunder, train, truck
object group	run	engine, horse, train, water
object group	scream	seagull, woman
object group	screech	owl, tire
object group	sing	whale, woman
object group	snore	dog, person
object group	sputter	engine, faucet
object group	squeak	door, mouse
object group	squeal	car brakes, pig
object group	wail	baby, siren
object group	whine	dog, drill, mosquito, vacuum cleaner
object group	whir	fan, helicopter
object group	whistle	bird, kettle, train, wind
action group	baby	babble, coo, cry, giggle, laugh, whimper

Kind	Group	Members
action group	bell	clang, clink, jingle, toll
action group	bird	chirp, flap, sing, squawk
action group	car	honk, skid
action group	cat	growl, hiss, meow, purr, scratch, yowl
action group	crowd	boo, cheer, clap, groan, laugh, whistle
action group	dog	bark, growl, pant, whimper
action group	door	bang, creak, slam, swing
action group	drum	beat, roll
action group	engine	idle, rev
action group	fire	crackle, roar
action group	glass	clink, shatter
action group	goat	bleat, scream
action group	goose	hiss, honk
action group	hen	cluck, squawk
action group	horse	gallop, neigh, snort, trot
action group	insect	buzz, chirp
action group	man	cough, laugh, shout, sneeze
action group	owl	hoot, screech
action group	person	clap, groan, hum, laugh, snore, whistle
action group	phone	buzz, chime
action group	pig	oink, squeal
action group	rain	patter, pour
action group	river	roar, trickle
action group	rooster	cluck, crow
action group	shower	drip, run
action group	siren	blip, wail
action group	thunder	boom, crack
action group	train	brake, chug, pass, screech
action group	water	drip, run, splash, trickle
action group	wave	crash, lap
action group	wind	blow, rustle
action group	woman	hum, laugh, sing, sneeze

Table E.3: Scenario pairs for the three inter-source relation fields. Each event is written as object / action; arrangements are written with *then* (sequence), *while* (concurrency), and *over* (foreground over background).

Field	Arrangement A	Arrangement B
temporal sequence	engine / idle then engine / rev	engine / rev then engine / idle
temporal sequence	dog / bark then man / laugh	man / laugh then dog / bark
temporal sequence	duck / quack then man / laugh	man / laugh then duck / quack
temporal sequence	bird / chirp then man / whistle	man / whistle then bird / chirp
temporal sequence	man / cough then toilet / flush	toilet / flush then man / cough
temporal sequence	engine / start then man / sing	man / sing then engine / start
temporal sequence	goat / bleat then man / laugh	man / laugh then goat / bleat
temporal sequence	engine / rev then man / cough	man / cough then engine / rev
temporal sequence	dog / bark then woman / laugh	woman / laugh then dog / bark
temporal sequence	crowd / cheer then man / laugh	man / laugh then crowd / cheer
temporal sequence	engine / run then man / whistle	man / whistle then engine / run
temporal sequence	water / run then woman / sing	woman / sing then water / run
temporal sequence	motorcycle / idle then motorcycle / rev	motorcycle / rev then motorcycle / idle
temporal sequence	cat / meow then man / sneeze	man / sneeze then cat / meow
temporal sequence	man / laugh then plastic / crinkle	plastic / crinkle then man / laugh
temporal sequence	woman / sneeze then woman / laugh	woman / laugh then woman / sneeze
temporal sequence	bee / buzz then man / whistle	man / whistle then bee / buzz
temporal sequence	baby / cry then man / cough	man / cough then baby / cry
temporal sequence	crowd / applaud then man / whistle	man / whistle then crowd / applaud
temporal sequence	keyboard / click then man / whistle	man / whistle then keyboard / click
temporal sequence	engine / run then horn / honk	horn / honk then engine / run
temporal sequence	train / approach then horn / blow	horn / blow then train / approach

Field	Arrangement A	Arrangement B
temporal sequence	man / whistle then horn / honk	horn / honk then man / whistle
temporal sequence	man / cough then man / whistle	man / whistle then man / cough
temporal sequence	telephone / ring then man / laugh	man / laugh then telephone / ring
temporal sequence	match / strike then fire / crackle	fire / crackle then match / strike
temporal sequence	cork / pop then liquid / pour	liquid / pour then cork / pop
temporal sequence	gun / fire then bird / flap	bird / flap then gun / fire
temporal sequence	doorbell / ring then dog / bark	dog / bark then doorbell / ring
temporal sequence	door / bang then man / whistle	man / whistle then door / bang
temporal sequence	baby / cry then woman / sing	woman / sing then baby / cry
temporal sequence	glass / shatter then dog / bark	dog / bark then glass / shatter
temporal sequence	door / slam then baby / cry	baby / cry then door / slam
temporal sequence	horn / honk then engine / rev	engine / rev then horn / honk
temporal sequence	duck / quack then water / splash	water / splash then duck / quack
temporal sequence	cat / meow then dog / bark	dog / bark then cat / meow
temporal sequence	sheep / bleat then dog / bark	dog / bark then sheep / bleat
temporal sequence	bell / ring then crowd / cheer	crowd / cheer then bell / ring
temporal sequence	person / whistle then dog / bark	dog / bark then person / whistle
temporal sequence	microwave / beep then man / laugh	man / laugh then microwave / beep
temporal sequence	kettle / whistle then man / sing	man / sing then kettle / whistle
temporal sequence	hammer / pound then drill / run	drill / run then hammer / pound
temporal sequence	saw / cut then hammer / pound	hammer / pound then saw / cut
temporal sequence	clock / chime then person / walk	person / walk then clock / chime
temporal sequence	train / honk then dog / bark	dog / bark then train / honk
temporal sequence	alarm / sound then person / walk	person / walk then alarm / sound
temporal sequence	car / honk then man / shout	man / shout then car / honk
temporal sequence	toilet / flush then door / close	door / close then toilet / flush
temporal sequence	frog / croak then owl / hoot	owl / hoot then frog / croak
temporal sequence	rooster / crow then dog / bark	dog / bark then rooster / crow
temporal concurrency	bird / chirp while man / whistle	bird / chirp then man / whistle
temporal concurrency	dog / bark while man / cough	dog / bark then man / cough
temporal concurrency	keyboard / click while man / cough	keyboard / click then man / cough
temporal concurrency	baby / cry while woman / sing	baby / cry then woman / sing
temporal concurrency	engine / run while man / whistle	engine / run then man / whistle
temporal concurrency	man / cough while plastic / crinkle	man / cough then plastic / crinkle
temporal concurrency	man / sing while water / run	man / sing then water / run
temporal concurrency	cat / meow while woman / laugh	cat / meow then woman / laugh
temporal concurrency	engine / rev while tire / squeal	engine / rev then tire / squeal
temporal concurrency	duck / quack while man / laugh	duck / quack then man / laugh
temporal concurrency	man / whistle while piano / play	man / whistle then piano / play
temporal concurrency	frog / croak while cricket / chirp	frog / croak then cricket / chirp
temporal concurrency	wiper / squeak while rain / fall	wiper / squeak then rain / fall
temporal concurrency	seagull / cry while wave / crash	seagull / cry then wave / crash
temporal concurrency	sheep / bleat while dog / bark	sheep / bleat then dog / bark
temporal concurrency	dog / bark while piano / play	dog / bark then piano / play
temporal concurrency	baby / cry while violin / play	baby / cry then violin / play
temporal concurrency	man / whistle while rain / fall	man / whistle then rain / fall
temporal concurrency	dog / bark while vacuum / run	dog / bark then vacuum / run
temporal concurrency	bird / chirp while water / run	bird / chirp then water / run
temporal concurrency	telephone / ring while guitar / play	telephone / ring then guitar / play
temporal concurrency	engine / idle while piano / play	engine / idle then piano / play
temporal concurrency	kettle / whistle while violin / play	kettle / whistle then violin / play
temporal concurrency	man / whistle while keyboard / click	man / whistle then keyboard / click
temporal concurrency	dog / bark while rain / fall	dog / bark then rain / fall
temporal concurrency	siren / wail while crowd / cheer	siren / wail then crowd / cheer
temporal concurrency	drill / run while hammer / pound	drill / run then hammer / pound
temporal concurrency	cat / meow while bird / chirp	cat / meow then bird / chirp
temporal concurrency	horse / trot while bird / chirp	horse / trot then bird / chirp
temporal concurrency	guitar / play while crowd / cheer	guitar / play then crowd / cheer
temporal concurrency	bee / buzz while bird / chirp	bee / buzz then bird / chirp
temporal concurrency	train / pass while bell / ring	train / pass then bell / ring
temporal concurrency	baby / cry while dog / bark	baby / cry then dog / bark
temporal concurrency	man / whistle while wind / blow	man / whistle then wind / blow
temporal concurrency	woman / sing while engine / run	woman / sing then engine / run

Field	Arrangement A	Arrangement B
temporal concurrency	person / clap while piano / play	person / clap then piano / play
temporal concurrency	dog / bark while thunder / rumble	dog / bark then thunder / rumble
temporal concurrency	bird / chirp while lawnmower / run	bird / chirp then lawnmower / run
temporal concurrency	man / whistle while printer / run	man / whistle then printer / run
temporal concurrency	crowd / cheer while piano / play	crowd / cheer then piano / play
temporal concurrency	duck / quack while water / run	duck / quack then water / run
temporal concurrency	phone / buzz while keyboard / click	phone / buzz then keyboard / click
temporal concurrency	siren / wail while dog / bark	siren / wail then dog / bark
temporal concurrency	engine / run while person / whistle	engine / run then person / whistle
temporal concurrency	rain / fall while thunder / rumble	rain / fall then thunder / rumble
temporal concurrency	bell / ring while crowd / cheer	bell / ring then crowd / cheer
temporal concurrency	vacuum / run while violin / play	vacuum / run then violin / play
temporal concurrency	airplane / fly while man / whistle	airplane / fly then man / whistle
temporal concurrency	motorcycle / rev while crowd / cheer	motorcycle / rev then crowd / cheer
temporal concurrency	baby / cry while water / run	baby / cry then water / run
foreground-background	bird / chirp over man / whistle	man / whistle over bird / chirp
foreground-background	engine / run over man / whistle	man / whistle over engine / run
foreground-background	dog / bark over man / whistle	man / whistle over dog / bark
foreground-background	man / sing over water / run	water / run over man / sing
foreground-background	insect / buzz over man / cough	man / cough over insect / buzz
foreground-background	baby / cry over person / snore	person / snore over baby / cry
foreground-background	bird / chirp over insect / buzz	insect / buzz over bird / chirp
foreground-background	man / whistle over pigeon / coo	pigeon / coo over man / whistle
foreground-background	bird / chirp over dog / bark	dog / bark over bird / chirp
foreground-background	bell / chime over person / clap	person / clap over bell / chime
foreground-background	cow / moo over tractor / run	tractor / run over cow / moo
foreground-background	typewriter / clack over clock / tick	clock / tick over typewriter / clack
foreground-background	goat / bleat over stream / flow	stream / flow over goat / bleat
foreground-background	owl / hoot over fire / crackle	fire / crackle over owl / hoot
foreground-background	man / whistle over guitar / play	guitar / play over man / whistle
foreground-background	dog / bark over car / honk	car / honk over dog / bark
foreground-background	baby / cry over piano / play	piano / play over baby / cry
foreground-background	bird / chirp over water / run	water / run over bird / chirp
foreground-background	crowd / cheer over violin / play	violin / play over crowd / cheer
foreground-background	telephone / ring over man / cough	man / cough over telephone / ring
foreground-background	dog / bark over baby / cry	baby / cry over dog / bark
foreground-background	siren / wail over man / laugh	man / laugh over siren / wail
foreground-background	rain / fall over man / sing	man / sing over rain / fall
foreground-background	piano / play over woman / sing	woman / sing over piano / play
foreground-background	crowd / cheer over guitar / play	guitar / play over crowd / cheer
foreground-background	duck / quack over water / splash	water / splash over duck / quack
foreground-background	engine / run over piano / play	piano / play over engine / run
foreground-background	hammer / pound over man / whistle	man / whistle over hammer / pound
foreground-background	bell / ring over crowd / cheer	crowd / cheer over bell / ring
foreground-background	vacuum / run over violin / play	violin / play over vacuum / run
foreground-background	cat / meow over man / whistle	man / whistle over cat / meow
foreground-background	train / pass over man / whistle	man / whistle over train / pass
foreground-background	thunder / rumble over rain / fall	rain / fall over thunder / rumble
foreground-background	keyboard / click over guitar / play	guitar / play over keyboard / click
foreground-background	horse / trot over bird / chirp	bird / chirp over horse / trot
foreground-background	motorcycle / rev over crowd / cheer	crowd / cheer over motorcycle / rev
foreground-background	guitar / play over man / sing	man / sing over guitar / play
foreground-background	water / splash over crowd / cheer	crowd / cheer over water / splash
foreground-background	drum / beat over crowd / cheer	crowd / cheer over drum / beat
foreground-background	dog / bark over lawnmower / run	lawnmower / run over dog / bark
foreground-background	owl / hoot over frog / croak	frog / croak over owl / hoot
foreground-background	person / whistle over dog / bark	dog / bark over person / whistle
foreground-background	alarm / sound over man / snore	man / snore over alarm / sound
foreground-background	kettle / whistle over man / whistle	man / whistle over kettle / whistle
foreground-background	saw / cut over hammer / pound	hammer / pound over saw / cut
foreground-background	bee / buzz over bird / chirp	bird / chirp over bee / buzz
foreground-background	sheep / bleat over dog / bark	dog / bark over sheep / bleat
foreground-background	airplane / fly over man / whistle	man / whistle over airplane / fly

Field	Arrangement A	Arrangement B
foreground-background	clock / tick over man / whistle	man / whistle over clock / tick
foreground-background	horn / honk over crowd / cheer	crowd / cheer over horn / honk

Table E.4: Anchor pool. Each anchor is an object–action event, listed with the fields whose value pairs it carries.

Object	Action	Fields
accordion	play	timbre
airplane	fly	distance, direction, loudness dynamics, timbre dynamics, distance dynamics
alarm	sound	pitch, duration, pitch dynamics, location, style
ambulance	wail	distance dynamics, direction dynamics
baby	cry	loudness, duration, distance, acoustic environment, location, style
basketball	bounce	style
bee	buzz	direction, distance dynamics, direction dynamics
bell	chime	pitch
bell	ring	count, pitch, loudness, timbre, duration, repetition, distance, distance dynamics, acoustic environment, location
bench	creak	material
bicycle	ring	direction dynamics
bird	chirp	count, loudness, repetition, distance, direction, direction dynamics, repetition dynamics, location, style
bird	sing	pitch, pitch dynamics
bowl	scrape	material
box	thud	material
bucket	hit	material
buzzer	sound	pitch
car	honk	direction dynamics, style
cat	meow	count, repetition, distance, distance dynamics, direction dynamics, location
cello	play	emotion
chair	scrape	material
clock	tick	repetition dynamics, style
cow	moo	count, loudness, direction dynamics
crow	caw	count, direction
crowd	applaud	duration
crowd	cheer	loudness, duration, distance, loudness dynamics, timbre dynamics, distance dynamics, acoustic environment, style
cup	clink	material
dish	clank	material
dog	bark	count, loudness, duration, repetition, distance, loudness dynamics, distance dynamics, direction dynamics, repetition dynamics, acoustic environment, location
dog	growl	direction dynamics
dog	whine	pitch dynamics
door	bang	material
door	slam	material, repetition
doorbell	ring	repetition, acoustic environment, style
drawer	slide	material
drum	beat	pitch, repetition, loudness dynamics, repetition dynamics, style
duck	quack	count, repetition, direction dynamics
engine	idle	timbre, timbre dynamics
engine	rev	pitch
engine	run	pitch, loudness, timbre, duration, distance, loudness dynamics, pitch dynamics, timbre dynamics, distance dynamics, direction dynamics, location
fan	hum	timbre, loudness dynamics, timbre dynamics
firework	bang	direction

Object	Action	Fields
floor	creak	material
flute	play	pitch
fly	buzz	direction dynamics
frog	croak	count
gate	slide	material
glass	shatter	style
goat	bleat	count
goose	honk	distance dynamics
guitar	play	timbre dynamics, emotion
harmonica	play	timbre
helicopter	fly	direction, loudness dynamics, distance dynamics, direction dynamics
helicopter	hover	direction, timbre dynamics
horn	blow	pitch
horn	honk	count, pitch, loudness, timbre, duration, repetition, distance, repetition dynamics, acoustic environment, style
horse	gallop	location
horse	neigh	direction dynamics
horse	trot	distance dynamics
insect	buzz	direction
jet	fly	direction, direction dynamics
kettle	whistle	pitch, loudness dynamics, pitch dynamics, timbre dynamics, style
keyboard	click	loudness, repetition dynamics
lawnmower	run	direction dynamics
machine	beep	repetition dynamics
man	scream	emotion
man	sing	pitch dynamics, emotion
man	walk	acoustic environment
motor	hum	pitch dynamics
motorcycle	idle	pitch dynamics
motorcycle	rev	count, timbre dynamics, distance dynamics, direction dynamics
owl	hoot	count, direction
⟨person⟩	cheer	human specifics
person	clap	repetition, repetition dynamics, acoustic environment
⟨person⟩	cough	human specifics
person	cough	location
⟨person⟩	cry	human specifics
⟨person⟩	grunt	human specifics
person	hum	direction dynamics, emotion
person	laugh	emotion
⟨person⟩	laugh	human specifics
person	run	distance dynamics
person	scream	emotion
⟨person⟩	scream	human specifics
⟨person⟩	shout	human specifics
person	shuffle	direction
⟨person⟩	sing	human specifics
⟨person⟩	sneeze	human specifics
person	sneeze	repetition
person	snore	distance
⟨person⟩	snore	human specifics
person	walk	distance dynamics
person	whistle	pitch, timbre, pitch dynamics, direction dynamics, location, emotion
piano	play	location, emotion
pipe	tap	material
plate	clank	material
rain	fall	loudness dynamics, timbre dynamics
seagull	cry	direction
sheep	bleat	count, direction dynamics
shutter	bang	material

Object	Action	Fields
siren	blare	pitch
siren	sound	pitch, pitch dynamics, distance dynamics, direction dynamics
siren	wail	pitch, duration, distance, loudness dynamics, timbre dynamics, location, style
spoon	clatter	material
surface	scratch	material
surface	tap	material
table	knock	material
telephone	ring	pitch, repetition, repetition dynamics, location, style
thunder	roar	repetition, direction
thunder	rumble	loudness, distance, direction, loudness dynamics, distance dynamics, style
traffic	hum	loudness dynamics
train	pass	loudness dynamics, timbre dynamics, style
train	rumble	distance dynamics, direction dynamics
train	whistle	distance dynamics, direction dynamics
tray	clatter	material
truck	honk	direction dynamics
truck	rumble	distance dynamics, direction dynamics
typewriter	clack	style
vacuum	run	timbre, timbre dynamics
vehicle	drive	loudness dynamics, distance dynamics
violin	play	timbre, loudness dynamics, acoustic environment, location, emotion
water	run	loudness dynamics
water	splash	loudness, duration
wind	blow	loudness, timbre, duration, loudness dynamics
woman	sing	emotion
woodpecker	peck	repetition, repetition dynamics

F CLAP, FAD, AND KAD MEASUREMENT

The CLAP score is computed with LAION-CLAP (Wu et al., 2023) (checkpoint 630k-audioset-fusion-best), each model scored as the mean over its 2,500 benchmark outputs and over the 953 clips it generates from the AudioCaps test captions, one caption per available test recording. FAD and KAD are computed with the kadtk toolkit (Chung et al., 2025) on PANNs Wavegram-Logmel-CNN14 embeddings, against the 953 real recordings, which enter Table F.1 as the reference row with both distances zero by construction. Figure F.1 plots the CLAP score on the benchmark outputs against the benchmark aggregate of Table 3; over the eleven models the Spearman correlation between the two is .18.

Table F.1: Benchmark aggregate (normalized gain) and CLAP, FAD, and KAD readings for the eleven models. Best value per column in bold, lower is better for FAD and KAD.

Model	Benchmark	CLAP (benchmark outputs)	CLAP (AudioCaps captions)	FAD	KAD
EzAudio	**.388**	.324	.517	**13.8**	**0.13**
TangoFlux	.384	.339	**.544**	19.9	0.59
Tango 2	.336	.301	.528	16.7	0.43
AudioX	.322	.354	.475	16.2	0.47
Stable Audio 3 SFX	.318	.371	.317	28.8	2.18
Stable Audio 3 Medium	.333	.390	.329	24.9	1.64
Stable Audio Open	.314	**.422**	.295	37.4	4.29
AudioGen	.292	.311	.467	14.0	0.21
MMAudio	.295	.346	.482	15.6	0.38
AudioLDM 2	.269	.318	.407	21.0	1.53
MAGNeT	.257	.310	.376	32.2	3.33
Real recordings (reference)	—	—	.498	0	0



Figure F.1: CLAP score on the benchmark outputs against the normalized gain of Table 3, one point per model.

G Aggregation Protocol and Stability

Every benchmark number in §5.3 averages three complete generation runs and, for the choice probes, permuted option orders: both orders for the two-option probes and four Latin-square orders for the *Dual-pair*, with the pair product taken within each order and the control averaged over the same orders. Table G.1 reports each model's normalized gain per run, after order averaging within the run. The three runs agree closely. The per-run totals stay within .026 of one another for every model, and averaging over option orders on its own moves model totals by at most .004 and changes two adjacent pairs, AudioX and Stable Audio 3 SFX and MMAudio and AudioGen, whose totals differ by .004 and .003. TangoFlux and EzAudio trade first place across runs. Prompt-sampling uncertainty is estimated with a bootstrap over prompts, 2,000 replicates resampled within field at the prompt or pair level with the control floors resampled jointly; Table G.2 reports the resulting intervals.

Table G.1: Normalized gain per generation run, after order averaging. Range = max - min across the three runs.

Model	**Run 1**	**Run 2**	**Run 3**	**Range**
EzAudio	.385	.384	.396	.013
TangoFlux	.387	.381	.384	.005
Tango 2	.338	.336	.334	.004
Stable Audio 3 Medium	.339	.330	.331	.009
AudioX	.330	.326	.311	.019
Stable Audio 3 SFX	.321	.317	.317	.004
Stable Audio Open	.317	.302	.322	.020
MMAudio	.298	.286	.302	.016
AudioGen	.282	.292	.303	.021
AudioLDM 2	.271	.264	.271	.007
MAGNeT	.268	.243	.260	.025

Table G.2: Normalized gain and *Match* with bootstrap 95% confidence intervals.

Model	Normalized gain	95% CI	Match	95% CI
EzAudio	.388	[.369, .408]	.840	[.830, .849]
TangoFlux	.384	[.364, .404]	.811	[.801, .821]
Tango 2	.336	[.315, .354]	.777	[.765, .788]
Stable Audio 3 Medium	.333	[.314, .352]	.713	[.701, .726]
AudioX	.322	[.304, .342]	.773	[.762, .784]
Stable Audio 3 SFX	.318	[.299, .337]	.742	[.730, .755]
Stable Audio Open	.314	[.294, .332]	.729	[.716, .742]
MMAudio	.295	[.277, .313]	.702	[.689, .714]
AudioGen	.292	[.274, .310]	.738	[.726, .749]
AudioLDM 2	.269	[.251, .285]	.680	[.666, .693]
MAGNeT	.257	[.239, .274]	.724	[.712, .736]

AGGREGATES WITHOUT THE CEILING

Table G.4 lists the raw per-field scores of all 25 fields, the judge's mean probability of the target answer, from which the normalized gains of Table 3 are computed. Table G.3 reports three aggregates that do not use the human ceiling: the normalized gain computed without the ceiling, $(S - F)/(1 - F)$, averaged over all 25 fields, with bootstrap 95% confidence intervals from the same resamples as Table G.2; a reliability-weighted variant, in which each field is weighted by the judge's reliability on that field, measured in the validation study as the leakage-scaled gain of the field's selected probe (the Judge reliability column of Table 3); and the raw mean of the 25 field scores. It is the normalized gain of §5.2 with the maximum score, 1, in place of the measured ceiling, so it is lower for every model and rests on distance and location as well; as audio-language judges improve, the ceilings move toward 1 and the two aggregates converge. The ordering is the same as under the ceiling except within the middle group.

Table G.3: Aggregates without the ceiling: normalized gain $(S-F)/(1-F)$ over the 25 fields with bootstrap 95% confidence intervals, its reliability-weighted variant, and the raw mean of the field scores.

Model	Without ceiling	95% CI	Reliability-weighted	Raw mean
TangoFlux	.322	[.307, .337]	.340	.410
EzAudio	.321	[.307, .336]	.338	.406
Tango 2	.288	[.272, .302]	.308	.378
Stable Audio 3 Medium	.277	[.263, .292]	.297	.367
Stable Audio 3 SFX	.270	[.254, .285]	.290	.361
AudioX	.269	[.254, .284]	.285	.361
Stable Audio Open	.263	[.247, .278]	.277	.355
AudioGen	.249	[.235, .263]	.266	.346
MMAudio	.246	[.232, .259]	.257	.340
AudioLDM 2	.229	[.215, .242]	.244	.323
MAGNeT	.226	[.211, .239]	.247	.324

H JUDGE ROBUSTNESS

To test whether the benchmark ranking depends on the choice of judge, we rescored the complete benchmark with the two alternate judges from the validation comparison, Qwen2.5-Omni and AudioFlamingo-Next, under the protocol of the primary Qwen3-Omni runs: the same frozen prompts and generated clips, the same probe questions, the same three generation runs, the same option-order averaging, and each judge's own Gaussian-noise floors. Absolute scores are not comparable across judges because the no-information floors differ, with a mean *Match* floor of .048 for Qwen3-Omni against .068 for Qwen2.5-Omni and .194 for AudioFlamingo-Next, so we compare rankings.

Model-level rankings are stable under judge replacement. The Spearman correlation between each alternate judge's ranking by the normalized gain computed without the ceiling (Appendix G; the

Table G.4: Raw per-field scores of the eleven models, the judge's mean probability of the target answer, with the reference columns of Table 3; three best models per row shaded. Distance and location are judge-limited on generated clips (§5.2) and have no ceiling. SA3, Stable Audio 3; SAO, Stable Audio Open.

Field	Probe	Reference				Models										
		Judge reliability	Control	Real audio	Ceiling	EzAudio	TangoFlux	Tango 2	AudioX	SA3 SFX	SA3 Medium	SAO	AudioGen	MMAudio	AudioLDM 2	MAGNeT
object	*Dual-pair*	.583	.005	.566	.974	.598	.604	.538	.575	.518	.497	.564	.507	**.672**	.377	.393
action	*Dual-pair*	.498	.012	.459	.987	.460	.465	.368	.387	.463	.417	.405	.356	**.468**	.357	.218
material	*Dual-pair*	.589	.008	.534	.655	.057	.132	.075	.104	.072	.105	.128	.122	**.183**	.069	.070
count	*Complete*	.786	.557	.811	.999	**.779**	.757	.738	.764	.752	.749	.731	.734	.734	.753	.763
human specifics	*Complete*	.893	.477	.906	.999	.775	.796	.791	.777	.760	.797	**.822**	.737	.806	.778	.747
pitch	*Dual-pair*	.621	.033	.587	.943	.229	.094	.096	.180	.158	.230	**.251**	.087	.094	.163	.045
loudness	*Dual-pair*	.655	.010	.610	.792	.124	.106	.055	**.129**	.103	.068	.086	.077	.045	.065	.079
timbre	*Dual-pair*	.669	.012	.605	.618	**.142**	.106	.065	.140	.097	.073	.084	.087	.066	.056	.042
duration	*Dual-pair*	.771	.015	.718	.832	**.170**	.089	.053	.119	.065	.055	.044	.055	.018	.143	.034
repetition	*Complete*	.819	.598	.863	.999	**.824**	.788	.792	.809	.810	.804	.774	.778	.792	.784	.705
distance	*Dual-pair*	.499	.013	.460	–	.031	.026	.012	.021	.022	.009	**.041**	.038	.031	.034	.010
direction	*Target*	.492	.237	.685	.768	.418	.407	.434	.428	.428	**.475**	.465	.348	.365	.381	.354
loudness dynamics	*Contrast*	.739	.130	.851	.866	**.361**	.284	.177	.173	.166	.167	.171	.160	.175	.194	.168
pitch dynamics	*Target*	.866	.011	.817	.936	.348	.439	.412	.384	.564	**.617**	.390	.344	.264	.244	.435
timbre dynamics	*Target*	.727	.213	.726	.845	.548	.514	**.550**	.489	.472	.471	.436	.492	.375	.413	.525
distance dynamics	*Target*	.907	.110	.907	.973	.688	.699	.737	.593	**.769**	.762	.660	.725	.457	.538	.700
direction dynamics	*Target*	.921	.213	.895	.989	.459	.512	**.515**	.402	.459	.391	.385	.459	.396	.367	.484
repetition dynamics	*Contrast*	.555	.135	.854	.452	**.231**	.175	.142	.205	.185	.202	.164	.151	.219	.137	.075
acoustic environment	*Contrast*	.863	.157	.906	.797	.327	.309	.305	.244	.255	.302	**.335**	.235	.334	.188	.211
location	*Dual-pair*	.565	.000	.556	–	.015	.035	**.057**	.010	.004	.003	.015	.002	.039	.002	.005
emotion	*Target*	.862	.013	.888	.975	.555	.551	.553	.575	.574	**.578**	.460	.448	.529	.544	.459
style	*Target*	.652	.383	.776	.983	.680	.646	.627	.618	.682	.698	.704	.640	.713	**.772**	.716
temporal sequence	*Target*	.645	.030	.787	.999	.741	**.965**	.767	.454	.332	.292	.312	.604	.311	.296	.493
temporal concurrency	*Contrast*	.716	.123	.776	.873	.324	**.457**	.388	.276	.157	.175	.174	.232	.159	.152	.225
foreground / background	*Contrast*	.637	.154	.798	.645	.261	**.290**	.207	.173	.164	.233	.268	.226	.257	.272	.156
Match (raw probability, all prompts)		.048	–	–		**.840**	.811	.777	.773	.742	.713	.729	.738	.702	.680	.724

ceilings are measured for the primary judge) and the primary judge's is .95 for Qwen2.5-Omni and .89 for AudioFlamingo-Next, between .87 and .93 on the raw and floor-subtracted variants, and .95 and .93 on the *Match* ranking. All three judges select the same top pair, TangoFlux and EzAudio, whose separation is statistically indistinguishable under the primary judge (§5.3), and the same bottom pair, AudioLDM 2 and MAGNeT. The largest displacements are AudioX, ranked sixth by the primary judge, eighth by Qwen2.5-Omni, and ninth by AudioFlamingo-Next, and Stable Audio Open, seventh by the primary judge and fourth by AudioFlamingo-Next; every other model moves at most two positions. Correlations over the 275 individual field-model cells are lower, Pearson .78 to .86, as expected from small per-cell samples and judge-specific field sensitivities. The full-factor protocol is also internally stable within each alternate judge: averaging over option orders moves model totals by .001 to .002 on average, and per-run totals vary by less than .010 for every model-judge combination.

I Protocol of the Human Rating Comparison

All correlations are computed on RELATE's official test split. For IS and OS we apply RELATE's listener screening, which leaves 1,020 and 1,110 rated test pairs.

The judge's whole-caption score is its probability of yes to the *Match* question, posed verbatim to each pair; its raw correlations are given in the text. The *Match* row of Table 2 is a random forest fitted from that probability and the 23 mention indicators described below, without the field scores, so that the three Ours rows share the fitting procedure and the indicators and differ only in which judge scores they use. CLAPScore is the cosine similarity between the caption and the clip under LAION-CLAP (Wu et al., 2023), 630k-audioset-best checkpoint, or under MS-CLAP (Elizalde et al., 2024), 2023 checkpoint. Human-CLAP (Takano et al., 2025) is LAION-CLAP fine-tuned on the listener ratings of RELATE's training split, run with its released checkpoint. ELSA (Suzuki et al., 2026) is run with its released code and checkpoints on the same pairs; its paper reports relevance on the generated clips of the split only.

For the field rows, each caption is annotated into the 25 fields and the judge answers one question for every field the caption mentions; each pair is then represented, field by field, by whether the caption mentions the field and by the judge's mean probability on its questions. Distance and location are left out: the human evaluation of §5.2 finds that the judge scores generated clips that raters confirm for these two fields at the level of the control, so the models use the remaining 23 fields. Models are fitted on RELATE's training split, hyperparameters are chosen on its validation split by Spearman correlation, and the test split is used once. Among ridge regression, decision trees, random forests, gradient boosting, support vector regression, and a multilayer perceptron, the random forest scored highest on the validation split and is the model reported. Uncertainty is a cluster bootstrap over captions, with the paired gain of the combined model over the whole-caption score computed on the same resampled captions on both sides. Field contributions are Shapley values over the 23 fields with the random forest's test R^2 as the value function, estimated from 200 random orderings of the fields, so they sum to the test R^2 of the reported model. The three largest values are object .089, action .052, and temporal sequence .052 on relevance (test R^2 .231), object .092, temporal sequence .034, and action .025 on presence (.166), and temporal sequence .048, object .028, and action .010 on order (.095). The random forest's permutation importance puts the same three fields first on every criterion and correlates with these Shapley values at $r \geq .96$.

J Human Evaluation Protocol

The pool is the base-seed output of EzAudio, TangoFlux and Tango 2 on the benchmark prompts. For target and complete fields a unit is one clip; for contrast and dual fields it is one minimal pair, of which a contrast unit contributes the clip with the higher judge score and a dual unit both clips. For each field the units were ranked by the judge's frozen score, the upper half kept, and ten drawn uniformly at random, at most one per prompt pair: 250 units, 340 clips in the first round. Five raters, all authors, rated every clip three times over: one rater rated all clips and the other four were assigned in balanced pairs so that each covers all 25 fields. Items were shown in an individual pseudorandom order after six unscored practice items; one unit per field was shown a second time, at least fifty

positions later, to measure intra-rater consistency, and only the first presentation enters the analysis. Raters saw no model name, no judge score and no indication of how a clip was sampled.

Each item asked the field probe in the judge's wording, with the option that the requested content is not clearly present, followed by a degree slider, the *Match* question with a 0-to-10 relevance slider, and an audio-quality slider; only the first answer enters the ceiling, and the sliders are kept for a separate correlation analysis. A clip is realized when at least two of its three raters give the key answer; the escape option counts as not realized and is never an error; both clips of a dual pair must pass. The ceiling of a field is the judge's mean score over its realized units, at the upper end of a bootstrap 95% interval over units, capped at 1. A field stops when that interval lies above its control; the pre-registered plan allowed up to two further rounds of ten units for a field that does not separate, and reports a field that never separates as judge-limited. After the first round, distance and location were reported as judge-limited without the further rounds the plan allowed. Distance does not separate from its control. Location does, but only narrowly, and three readings say the judge is not detecting it: it scores the confirmed and the unconfirmed units alike, both far below its score on real recordings that state a location (Table J.1), and the raters noted repeatedly that the property was hard to judge. We therefore report location as judge-limited as well; more units would not change either reading. Duration and material, with two confirmed units each, received a second round of ten units drawn from the same pool in the same random order, rated by two of the authors rather than the full panel, with a unit confirmed only when both gave the key answer, a stricter rule than the two-of-three rule of the first round; both departures from the pre-registered plan are recorded in a dated note, and the ceilings of these two fields pool the units of both rounds. Raters gave the same answer on 73% of the repeated items in the first round, and the two raters of the second round gave the same answer on 28 of its 40 clips. Table J.1 lists the result for every field, next to the judge's mean score on real recordings that state the field (the Real audio column of Table 3).

Table J.1: Human evaluation results per field. Mean is the judge's raw mean score on the confirmed units; Real audio is the same probe's mean score on the validation study's real recordings, repeated from Table 3.

Field	**Confirmed**	**Mean**	**Real audio**	**95% interval**	**Ceiling**	**Control**	**Outcome**
object	9/10	.918	.566	[.850, .974]	.974	.005	separated
action	8/10	.830	.459	[.607, .987]	.987	.012	separated
material	5/20	.307	.534	[.076, .655]	.655	.008	separated
count	10/10	.995	.811	[.992, .999]	.999	.557	separated
human specifics	10/10	.997	.906	[.995, .999]	.999	.477	separated
pitch	3/10	.808	.587	[.704, .943]	.943	.033	separated
loudness	3/10	.304	.610	[.020, .792]	.792	.010	separated
timbre	4/10	.286	.605	[.085, .618]	.618	.012	separated
duration	5/20	.486	.718	[.140, .832]	.832	.015	separated
repetition	8/10	.997	.863	[.994, .999]	.999	.598	separated
distance	4/10	.011	.460	[.003, .022]	.022	.013	judge-limited
direction	6/10	.642	.685	[.509, .768]	.768	.237	separated
loudness dynamics	6/10	.665	.851	[.458, .866]	.866	.130	separated
pitch dynamics	5/10	.882	.817	[.816, .936]	.936	.011	separated
timbre dynamics	5/10	.654	.726	[.492, .845]	.845	.213	separated
distance dynamics	9/10	.953	.907	[.931, .973]	.973	.110	separated
direction dynamics	8/10	.906	.895	[.795, .989]	.989	.213	separated
repetition dynamics	5/10	.294	.854	[.135, .452]	.452	.135	separated
acoustic environment	10/10	.604	.906	[.413, .797]	.797	.157	separated
location	3/10	.043	.556	[.003, .100]	.100	.000	judge-limited
emotion	9/10	.937	.888	[.890, .975]	.975	.013	separated
style	7/10	.949	.776	[.911, .983]	.983	.383	separated
temporal sequence	9/10	.999	.787	[.998, .999]	.999	.030	separated
temporal concurrency	8/10	.728	.776	[.577, .873]	.873	.123	separated
foreground / background	9/10	.548	.798	[.449, .645]	.645	.154	separated