This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Omni2Sound: Towards Unified Video-Text-to-Audio Generation

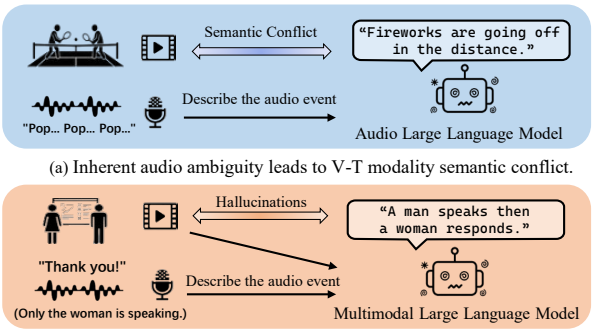
Yusheng Dai[2,3], Zehua Chen[1,3†], Yuxuan Jiang[1,3],
Qiuhong Ke[2], Jianfei Cai[2], Jun Zhu[1,3†]
[1] Tsinghua University, Beijing, China [2] Monash University, Melbourne, Australia
[3] Shengshu AI, Beijing, China

Abstract

*Training a unified model integrating video-to-audio (V2A), text-to-audio (T2A), and joint video-text-to-audio (VT2A) generation offers significant application flexibility, yet faces two unexplored foundational challenges: (1) the scarcity of high-quality audio captions with tight V-A-T alignment, leading to severe semantic conflict between multimodal conditions, and (2) cross-task and intra-task competition, manifesting as an adverse V2A-T2A performance trade-off and modality bias in the VT2A task. First, to address data scarcity, we introduce **SoundAtlas**, a large-scale dataset (470k pairs) that significantly outperforms existing benchmarks and even human experts in quality. Powered by a novel agentic pipeline, it integrates Vision-to-Language Compression to mitigate visual bias of MLLMs, a Junior-Senior Agent Handoff for a 5× cost reduction, and rigorous Post-hoc Filtering to ensure fidelity. Consequently, SoundAtlas delivers semantically rich and temporally detailed captions with tight V-A-T alignment. Second, we propose **Omni2Sound**, a unified VT2A diffusion model supporting flexible input modalities. To resolve the inherent cross-task and intra-task competition, we design a three-stage multi-task progressive training schedule that converts cross-task competition into joint optimization and mitigates modality bias in the VT2A task, maintaining both audio-visual alignment and off-screen audio generation faithfulness. Finally, we construct **VGGSound-Omni**, a comprehensive benchmark for unified evaluation, including challenging off-screen tracks. With a standard DiT backbone, Omni2Sound achieves unified SOTA performance across all three tasks within a single model, demonstrating strong generalization across benchmarks with heterogeneous input conditions.*

1. Introduction

Early audio generation models typically rely on unimodal conditioning. Text-to-Audio (T2A) [1–6] offers strong semantic fidelity and generalization but lacks dense temporal control. Conversely, Video-to-Audio (V2A) [7–10] ensures fine-grained temporal synchronization with video, yet suffers from weak reasoning in complex scenes and unfaithful generation (e.g., unexpected music or speech) [11, 12]. To address this, recent Video-Text-to-Audio (VT2A) methods [11, 13–16] jointly condition on video and text. While VT2A achieves both strong semantic understanding and temporal alignment, its reliance on simultaneous inputs constrains its applicability [16]. Crucially, most VT2A systems lack robustness [11, 13, 14], degrading sharply under missing-modality conditions (video-only or text-only).



(a) Inherent audio ambiguity leads to V-T modality semantic conflict.

(b) Visual modality bias in native MLLMs leads to hallucinations.

Figure 1. Challenges in scaling high-quality audio captions.

These constraints motivate a unified framework natively supporting VT2A, V2A, and T2A within a single model. This unified paradigm aligns with the AIGC shift, eliminating the redundant architectures and deployment complexity of hard-switching between specialized models. Recent work has begun to advance this unified approach. MMAudio [15] introduces a multimodal joint training framework to improve V2A generation, optionally conditioning on text using large-scale text–audio pairs. Moreover, AudioX [16] enhances flexibility by supporting broader modality combinations. Despite this progress, two challenges in the unified VT2A framework remain underexplored. [1]

First, there is a scarcity of high-quality audio captions that are well-aligned with both audio and video

† Corresponding author.

[1] https://omni2sound.github.io

cues. Most unified or specialized VT2A studies create their (V, T, A) training triplets by pairing videos (V) and their audio (A) with captions (T) generated solely from the audio [13, 16]. However, this approach introduces severe semantic conflict in the multimodal training data (see Figure 1): a frequent mismatch between the visual content and the (audio-only) text caption. This conflict is rooted in the audio modality's inherent ambiguity (e.g., a tennis hit vs. distant fireworks, or car engine noise vs. an electric drill). This fundamental ambiguity is then exacerbated by the limited capabilities of earlier audio-language models, which are prone to severe hallucinations (e.g., omissions and mislabels) [17]. In our preliminary experiments, we found these modality conflicts caused by mismatches between V-T conditions directly lead to unstable convergence and a significant degradation in audio faithfulness. Unfortunately, there is still a lack of high-quality V-T-A triples for unified VT2A models training, as we further discuss in Section 2.

Second, two critical types of task competition within unified VT2A frameworks remain underexplored. (1) Cross-Task Competition. Prior work, notably MMAudio [15], established that incorporating T-A pairs enhances the generalization and quality of V2A generation. However, training a unified model to excel at both V2A and T2A presents a significant challenge: as shown in our preliminary experiment (Table 5), this joint training introduces a severe T2A-V2A adverse trade-off, rooted in the heterogeneity between text and video modalities. Prioritizing one task during training consistently degrades the performance of the other, indicating a zero-sum optimization dynamic. (2) Intra-Task Competition. We also observe competition within the VT2A task itself. This competition manifests as a modality bias during generation process that undermines cross-conditional consistency, revealing two key failure modes: a bias towards text leads to poor A-V synchronization (Table 6), while a bias towards video exhibits low text-audio faithfulness in off-screen generation scenarios (Table 7).

To address data scarcity, we first introduce ***SoundAtlas*** in Section 3, a large-scale, agent-generated multimodal audio-caption dataset. It augments the two largest audio datasets, VGGSound [18] and AudioSet [19], providing semantically rich and temporally detailed captions that even surpass human-expert quality (Table 2). Built on current advanced multimodal foundation models (Gemini-2.5 Pro [20] and Qwen-2.5-VL [21]), we develop a multi-turn, agentic annotation pipeline featuring a junior–senior agent handoff, vision-to-language compression, and post-hoc hallucination filtering. This pipeline delivers cost-controlled annotations while maintaining tight visual–audio–text (V–A–T) alignment and a markedly higher text-audio faithfulness than prior datasets. Interestingly, we find its quality is high enough to even correct human annotation errors in VGGSounder [22].

Building on this dataset, we propose ***Omni2Sound*** in Section 4, a diffusion-based unified model supporting flexible input modalities while maintaining both audio-visual synchronization and high-fidelity generation. To address cross-task and intra-task competition, we introduce a three-stage progressive training schedule that departs from naive joint training. First, a *large-scale T2A pretraining stage* establishes a robust generative prior, enabling minimal high-quality T2A replay in the subsequent stage to prevent catastrophic forgetting. Subsequently, our *Multi-task Interleaved Training* integrates V2A and T2A tasks with high-quality VT2A triplets. Our central insight is that this VT2A data serves as a semantic bridge: by aligning the heterogeneous feature spaces of video and text, it effectively converts zero-sum cross-task competition into a cooperative optimization dynamic, thereby mitigating training resource contention. To resolve the intra-task competition, our third stage employs a *decoupled Robustness Training*. We utilize two synergistic augmentations to balance cross-modal reliance: *Text Dropout* penalizes text bias to enhance A-V synchronization, while *Off-screen Synthesis* counteracts video bias to ensure textual faithfulness. This decoupled approach rectifies key failure modes, maintaining high-fidelity generation even in challenging, asymmetric input scenarios.

Finally, we construct ***VGGSound-Omni*** in Section 5, the first comprehensive benchmark to establish a unified evaluation standard for VT2A, V2A, and T2A. It provides high-quality, human-verified annotations for all three tasks and introduces a challenging off-screen audio generation track. As a result, with a vanilla DiT [23] backbone, Omni2Sound achieves unified state-of-the-art performance across all three tasks against both unified and specialized models, showing high-fidelity audio quality, tight audio-visual synchronization, and excellent generation faithfulness.

2. Related Works

Audio Caption Dataset. Human-annotated benchmarks like *AudioCaps* [24] (46k) and *Clotho* [25] (5k) offer high-quality alignment, but their limited scale, high cost, and lack of detail make them unsuitable for training modern, large-scale models. To address data scarcity, automated pipelines such as *WavCaps* [26] use LLMs to refine noisy web metadata (400k captions), and *AudioSetCaps* [27] uses ALMs+LLMs to extract and aggregate details from audio, speech, and music, significantly increasing data volume. However, due to the inherent ambiguity of the audio modality, these audio-only

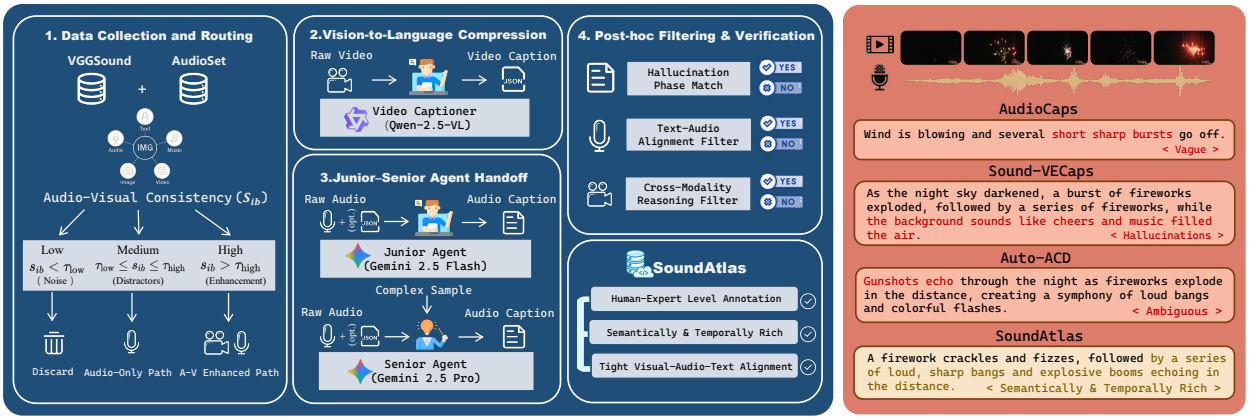


Figure 2. Data Construction Pipeline of SoundAtlas (Left). Comparison against SOTA baselines and human annotations (Right) .

methods suffer from high hallucination rates and can introduce cross-modal conflicts that destabilize VT2A training. Visually-enhanced (VE) annotation pipelines like *Auto-ACD* [28] and *Sound-VECaps* [29] leverage visual cues as cross-modal constraints. Despite their promise, existing VE pipelines adopt a separate-then-fuse design: unimodal models extract separate textual cues (e.g., image captions, audio tags), which are then merged by a final LLM. This is suboptimal, as the LLM operates on lossy textual representations rather than raw modalities, causing unimodal hallucinations to accumulate and amplify. Using native end-to-end multimodal models (e.g., *Gemini* [20]) appears to be a natural alternative, yet as we show in Section 3, this approach faces prohibitive costs and a pervasive visual bias that prevents audio-centric captioning. *There remains a lack of a large-scale, high-quality visual–audio–text (V–A–T) aligned audio caption dataset suitable for training unified VT2A models.*

Unified Audio Generation Model. The audio generation paradigm is shifting towards unified, omni-modal frameworks. This trend is initiated by *MMAudio* [15], which first integrates V2A and T2A but remains fundamentally V2A-centric, using T-A pairs merely as augmentation rather than optimizing T2A as a co-equal task. Subsequent works such as *AudioX* [16] and *AudioGen-Omni* [30] expand the scope to more flexible modality combinations, yet often rely on brute-force data scaling (e.g., *AudioX* with over 9 million samples) without achieving commensurate performance gains. Critically, these early models [15, 16, 30] largely overlook the inherent cross-task competition that arises from co-training diverse sub-tasks. *UniFlow-Audio* [31] is the first to systematically address this issue by categorizing tasks into Time-Aligned (TA) and Non-Time-Aligned (NTA) classes and analyzing their competitive dynamics. However, the analysis remains coarse-grained, without investigating the finer-grained competition within the TA category (i.e., V2A vs. T2A). Moreover, the challenging case of joint cross-modal generation (VT2A) remains unaddressed. *Consequently, a fundamental study on task competitive dynamics within a unified VT2A framework remains absent.*

3. SoundAtlas: V-A-T Data Construction

Existing automated audio caption datasets often suffer from severe visual-audio-text (V-A-T) misalignment and high hallucination rates due to the limited capability of early ALMs [27–29]. While recent native multimodal foundation models such as Gemini 2.5 [20, 32, 33] offer strong capabilities, we find that naively processing raw video-audio pairs is suboptimal for constructing audio caption datasets. Specifically, it incurs prohibitive costs (approx. $10,275 per 1M samples; see Appendix A) and introduces inherent visual bias, where models hallucinate auditory labels for visually suggested but non-existent events, as shown in Figure 1.

To address these challenges, we introduce SoundAtlas, constructed through a cost-effective, multi-turn agentic annotation pipeline. As illustrated in Figure 2, our pipeline integrates **vision-to-language compression** to mitigate visual bias, a **junior–senior agent handoff** to optimize cost-efficiency, and rigorous **post-hoc filtering** to ensure annotation fidelity. Full prompt instructions are provided in Appendix B.

A-V Consistency Routing. We first apply A-V Consistency Routing to raw videos from AudioSet [19] and VGGSound [18]. The key observation is that visual cues are reliable for high-consistency A-V clips but act as distractors in low-consistency ones, as shown in Figure 2. We classify samples by their ImageBind alignment score (s_{ib}) using thresholds $\tau_{\text{low}} = 0.20$ and $\tau_{\text{high}} = 0.30$: (i) High-consistency samples ($s_{ib} > \tau_{\text{high}}$) enter the *A-V Enhanced Path*; (ii) Medium-consistency samples

Table 1. Semantic Faithfulness (CLAP Score) of Different Data Construct Pipelines on AudioSet and VGGSound.

Method	AudioSet		VGGSound	
	LA-CLAP ↑	MS-CLAP ↑	LA-CLAP ↑	MS-CLAP ↑
AudioSetCaps [27]	0.330	0.397	0.351	0.421
Sound-VECaps [29]	0.370	0.425	-	-
Auto-ACD [28]	0.396	0.437	0.409	0.457
SoundAtlas (Ours)	**0.447**	**0.485**	**0.461**	**0.502**

Table 2. Caption quality comparison via MLLM-as-a-judge and human evaluation, reporting the Mean Win Rate for Semantic (MWR-S) and Temporal (MWR-T) alignment. Human-Expert refers to the human-annotated captions from AudioCaps [24].

Method	MLLM Evaluation		Human Evaluation	
	MWR-S ↑	MWR-T ↑	MWR-S ↑	MWR-T ↑
Auto-ACD [28]	0.39	0.41	0.31	0.26
Human-Expert [24]	0.36	0.51	0.46	0.55
SoundAtlas (Ours)	**0.75**	**0.58**	**0.71**	**0.69**

($\tau_{\text{low}} \leq s_{ib} \leq \tau_{\text{high}}$) are routed to the *Audio-Only Path* to prevent visual hallucinations; and (iii) Noise ($s_{ib} < \tau_{\text{low}}$) is discarded.

Vision-to-Language Compression. This step implements our key insight: vision should be treated as a contextual constraint, not a primary input. We find that compressing the visual stream into a textual representation (c_v) effectively addresses both challenges identified above. First, it reduces cost by replacing the prohibitively expensive raw video input ($V + A$) with a cost-effective text-audio prompt ($c_v + A$). Second, it mitigates cross-modal hallucinations by removing direct visual bias and providing only semantic context (e.g., "A man and a woman are standing...") rather than the raw visual stream. Concretely, for samples V routed to the *A-V Enhanced Path*, we use Qwen-2.5-VL [21] to analyze the video V (without its audio A) and generate the textual representation $c_v = \text{Qwen}(V)$. For samples on the *Audio-Only Path*, the visual context is set to null.

Junior–Senior Agent Handoff. All samples then enter our handoff pipeline. Each sample is first assigned to the Junior agent, G_{junior} (Gemini 2.5 Flash), which receives the audio A and the optional visual context c_v, producing a caption $c_a = G_{\text{junior}}(A, c_v)$. This caption c_a is flagged if it (i) triggers complexity criteria (text-based heuristics to identify complex audio scenes), (ii) contains high-frequency hallucination phrases, or (iii) fails a differentiated CLAP [34] check, $\text{CLAP}(c_a, A) < \tau_{clap}$, where τ_{clap} is 0.35 for general audio and 0.15 for music. Flagged samples are escalated to the Senior agent, G_{senior} (Gemini 2.5 Pro). To control costs, the reasoning output of the Senior agent is limited to 128 tokens, yielding a more precise caption.

Post-hoc Filtering and Verification. Finally, all generated captions c_a undergo two-stage verification. First, a CLAP (T-A) filtering model [34] ensures high Text-Audio faithfulness; captions where $\text{CLAP}(c_a, A) < \tau_{verify}$ are discarded. Second, for captions from the *A-V Enhanced Path* ($c_v \neq \emptyset$), an A-V-T Verifier, V_{AVT}, checks that c_a is a plausible acoustic description given c_v. Captions that pass all filters are accepted into the final dataset $\mathcal{D}_{\text{SoundAtlas (Ours)}}$, which augments VGGSound [18] and AudioSet [19] with human-expert-level audio captions.

3.1. Comparison with Existing Pipeline

We compare SoundAtlas against other automated pipelines [27–29] on high audio-visual consistency subsets sourced from AudioSet and VGGSound, where ImageBind score $s_{ib} > 0.30$. As shown in Table 1, SoundAtlas significantly outperforms all competitors on both LA-CLAP and MS-CLAP scores, demonstrating superior text-audio alignment. Additionally, we conduct a fine-grained MLLM-as-a-judge (Gemini 3.0 Pro [20]) evaluation on the intersection of AudioCaps and all compared datasets [24]. As shown in Table 2, SoundAtlas achieves a substantially higher mean win rate in semantic alignment (MWR-S) and temporal alignment (MWR-T) than both the strongest baseline (Auto-ACD) and the Human-Expert annotations, across both semantic and temporal alignment. To mitigate potential evaluation bias, a follow-up human validation study is conducted, further corroborating our results (details in Appendix Section C). As illustrated in Figure 2 (right), SoundAtlas demonstrates clear superiority over existing automated datasets, characterized by its richer semantic content and explicit temporal ordering.

4. Omni2Sound: Unified VT2A Generation

Building on SoundAtlas, we propose Omni2Sound, a Diffusion-based unified VT2A model supporting collaborative (VT2A) and unimodal (V2A, T2A) control.

4.1. Foundation Model Architecture

We adhere to a principle of simplicity and scalability, adopting a standard Diffusion Transformer (DiT) backbone [23] conditioned on latent features from a pretrained audio VAE [35]. As shown in Figure 3, the backbone is conditioned on multimodal inputs using a decoupled injection approach, which is separated into two distinct branches: (1) **Semantic Branch (*What*)** and (2) **Temporal Branch (*When*)**. To capture global semantic context, we concatenate text embeddings from Flan-T5 [36] (F_t) and visual features from CLIP [37] (F_v, sampled at 8 fps) along the temporal dimension, which are then injected via cross-attention layers. Crucially, this

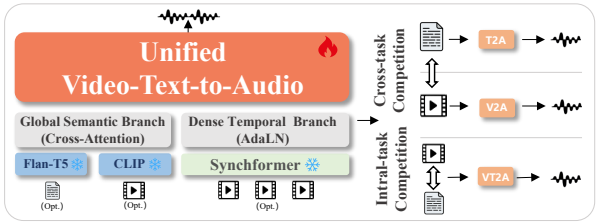


Figure 3. Overview of our unified VT2A framework, which integrates global semantics and temporal alignment, supporting flexible T2A, V2A, and VT2A generation.

design allows for flexible unimodal generation (V2A or T2A) by simply omitting the absent modality without requiring padding constraints. For the Temporal Branch, to ensure fine-grained synchronization, we follow [15] to utilize a Synchformer [38] to extract dense visual-temporal features (F_s) and then inject it globally via Adaptive Layer Normalization (AdaLN).

This decoupled architecture effectively (1) achieves the flexibility of multi-condition frameworks like AudioX [16], supporting extensible conditions without architectural modification; and (2) maintains precise temporal alignment comparable to MMAudio [15] (powered by its well-designed MM-DiT architecture).

4.2. Three-stage Progressive Multi-task Training

As established in Section 1, naive joint training suffers from both cross-task and intra-task competition. To address these challenges, we design a three-stage progressive training schedule.

Stage 1: Large-scale T2A Pretraining. We first pretrain the model on large-scale text-audio pairs without quality filtering. Following recent advances in latent generative modeling [3, 23, 39], our DiT backbone (Section 4.1) learns to denoise a noisy latent z_t at timestep t, conditioned on text embeddings H_c. The model ϵ_θ is optimized with the standard L2 loss:

$$L = \mathbb{E}_{t,z_t,\epsilon}\|\epsilon - \epsilon_\theta(z_t, t, H_c)\|^2$$

This pretraining serves two purposes. First, it establishes a robust generative prior before introducing heterogeneous video conditions. Second, it enables a significantly lower T2A sampling frequency in the subsequent stage without catastrophic forgetting, thereby reducing resource contention.

Stage 2: Multi-task Interleaved Training. This stage addresses cross-task competition through interleaved task sampling. At each step, a single task $s \in \{V2A, T2A, VT2A\}$ is sampled from a categorical distribution Cat(π), and a minibatch is drawn exclusively from its dataset D_s for a single-task gradient update. This avoids within-batch loss mixing and stabilizes optimization. Our ablations (Table 5) reveal two key findings: (i) The VT2A task serves as a critical bridge that mitigates the adverse V2A-T2A trade-off, enabling their simultaneous optimization rather than zero-sum competition. (ii) With this bridge in place, a low T2A sampling frequency (e.g., $\pi_{T2A} = 0.1$) on high-quality data suffices to prevent catastrophic forgetting. Together, these findings allow Stage 2 to focus primarily on video-conditioned tasks (V2A and VT2A), using T2A only sparingly to maintain its generative prior.

Stage 3: Intra-Task Resolution via Robustness Training. While Stage 2 resolves cross-task competition, intra-task competition (modality bias) persists, particularly for challenging scenarios such as off-screen generation. We therefore introduce a final robustness training stage. Crucially, this stage is decoupled from Stage 2: as shown in Table 6, introducing robustness augmentations prematurely destabilizes the multi-task optimization, whereas applying them after convergence enhances cross-modal consistency without compromising generative quality.

This stage employs two complementary augmentations to enforce balanced reliance on both modalities: (i) *Text Dropout.* We randomly drop tokens from the text prompt, creating ambiguity that compels the model to attend more to the visual stream. This counteracts text bias and strengthens audio-visual synchronization. (ii) *Off-screen Synthesis.* We incorporate off-screen audio samples and augment the text prompt to describe them, producing training pairs where the audio content is absent from the video. This counteracts video bias and improves textual faithfulness for off-screen audio generation.

5. VGGSound-Omni: Unified Evaluation

A key challenge in evaluating unified Video-Text-to-Audio (VT2A) models is the absence of a comprehensive benchmark. The VGGSound test set [18] provides only sparse event labels and lacks detailed captions. Although VGGSounder [22] improved upon this by correcting and introducing modality labels (e.g., A, V, AV) for fidelity evaluation, it still does not provide human-expert-level captions. To bridge this gap, we construct VGGSound-Omni, a multi-track benchmark derived from the VGGSound test set for evaluating both standard unified and specialized off-screen VT2A generation.

VGGSound-Omni Construction. We first establish a high-fidelity, human-level caption set covering all 14,000+ videos as the primary evaluation track. Ini-

Table 3. Comparison on VGGSound-Omni benchmark: Omni2Sound against SOTA models on T2A, V2A, and VT2A tasks. The *w/ Video-LLaMA caps* row evaluates Omni2Sound's generalization to unseen captions generated by Video-LLaMA [40].

Task	Method	Distribution Matching				Audio Quality		Modality Alignment		
		KL↓	FD↓	FAD↓	FD_{PaSST}↓	PQ↑	IS↑	DS↓	IB↑	MS-CLAP↑
T2A	AudioX [16]	1.68	9.04	1.42	109.94	6.37	15.15	-	-	0.49
	MMAudio [15]	1.92	8.62	1.63	101.66	5.84	14.30	-	-	0.50
	Omni2Sound (ours)	**1.53**	**4.61**	**1.01**	**60.38**	**6.52**	**16.41**	-	-	**0.53**
	w/ Video-LLaMA caps	1.60	6.92	1.23	83.91	6.38	16.01	-	-	0.51
V2A	V-AURA [41]	2.28	16.43	2.34	245.25	5.74	10.82	0.69	0.28	0.32
	Frieren [8]	2.73	12.13	1.23	123.75	5.82	11.32	0.86	0.21	0.31
	AudioX [16]	2.96	12.73	1.42	121.82	**6.17**	13.34	1.22	0.24	0.34
	MMAudio [15]	2.11	5.65	0.81	69.33	5.72	11.85	0.48	0.28	0.43
	Omni2Sound (ours)	**2.04**	**3.41**	**0.51**	**50.19**	6.15	**16.18**	**0.47**	**0.35**	**0.44**
VT2A	ThinkSound (w/o. CoT) [14]	1.60	7.41	1.10	116.08	**6.21**	11.73	0.53	0.26	0.43
	HunyuanVideo-Foley [13]	1.74	10.02	2.36	100.53	6.18	11.58	0.57	0.32	0.45
	AudioX [16]	1.59	8.29	1.24	103.37	6.17	14.94	1.23	0.26	0.49
	MMAudio [15]	1.63	5.28	0.91	68.44	5.84	13.44	**0.49**	0.29	0.49
	Omni2Sound (ours)	**1.35**	**2.95**	**0.53**	**48.20**	6.21	**15.79**	0.49	**0.34**	**0.52**
	w/ Video-LLaMA caps	1.56	3.37	0.66	53.73	6.11	15.74	0.50	0.34	0.49

tial captions are generated using our agentic pipeline (Section 3) and then systematically validate through an AI-assisted verification workflow: GPT-5 [42] served as an auditor, checking whether the captions semantically covered all "A" and "AV" labels from VGGSounder [22]. Samples flagged with mismatches are routed for targeted human verification. During this manual audit, we find that most flagged discrepancies stemmed from annotation errors in the VGGSounder data itself (e.g., label redundancy and errors caused by visual interference). After correcting these errors, we establish the final, human-verified captions as the definitive ground truth (GT) for all three tasks (VT2A, V2A, and T2A).

Complementing the primary set, we curate a challenging off-screen track (1,000+ items) from two sources: (i) *Natural events*, filtered from VGGSound for low A-V correspondence (via IB-Score [43] and Desync-Score [15]) while excluding background speech; and (ii) *Synthetic music*, formed by mixing aligned background clips from MusicCaps [44]. More details are provided in Appendix D.

6. Experiments

6.1. Experiment Settings

Datasets. For T2A backbone pre-training, we use a large-scale corpus comprising the train set of audio datasets such as AudioCaps [24], WavCaps [26], Clotho [25], AudioSet [19], VGGSound [18], FSD50k [45], as well as music datasets including MSD [46] and FMA [47]. To maintain consistency, all audio is segmented into 10-second clips and resampled at 16 kHz. Following this, the model is trained for unified VT2A tasks using our proposed SoundAtlas (Section 3) and a high-quality, PQ-score-filtered T-A subset derived from the aforementioned pre-training corpus. More details are provided in Appendix Section G. For evaluation, we compare Omni2Sound with SOTA models on three benchmarks: our proposed VGGSound-Omni (Section 5), Kling-Audio-Eval [30], AudioCaps test set [24] and AudioAtlas [48]. We ensure that these evaluation benchmarks are strictly disjoint from all data used in our training stages to prevent potential data leakage.

Evaluation Metrics. We implement our objective evaluation using the standardized AV-benchmark toolkit [15] on 8-second clips, following previous work [15]. We assess quality across four critical dimensions [2]. For **Distribution Matching**, we measure feature similarity between generated and ground-truth audio using Fréchet Distance (FAD [49], FD_{PaSST} [50], FD [51]) and Kullback-Leibler divergence (KL, KL_{PaSST}). **Audio Quality** is assessed via Inception Scores (IS [52], IS_{PaSST}) and Production Quality (PQ [53]) for aesthetics. **Semantic Alignment** evaluates text-audio consistency (CLAP [34], MS-CLAP [54]) and video-audio alignment (IB [43]). Finally, **Temporal Alignment** is measured using the Desynchronization Score (DS) predicted by Synchformer [55]. Detailed metric definitions and calculations are provided in the Appendix.

6.2. Main Results

Evaluation on VGGSound-Omni. We present our main results on VGGSound-Omni benchmark in Table 3. To ensure a fair comparison, all baseline models are re-evaluated using their official checkpoints and the standardized AV-benchmark [15], using the same video and text conditions. The results demonstrate that Omni2Sound achieves state-of-the-art performance across all three unified tasks (T2A, V2A, and VT2A) compared to both previous unified VT2A models (Au-

Table 4. Comparison on the Kling-Audio-Eval: Omni2Sound against SOTA models on T2A, V2A, and VT2A tasks.

Task	Method	Distribution Matching				Audio Quality		Modality Alignment		
		KL↓	FD↓	FAD↓	FD_{PaSST}↓	PQ↑	IS↑	DS↓	IB↑	LA-CLAP↑
T2A	AudioX [16]	2.73	19.43	3.32	171.60	5.98	**12.15**	-	-	0.28
	MMAudio [15]	2.54	**11.25**	5.07	**142.71**	5.54	9.28	-	-	0.28
	Omni2Sound (ours)	**2.36**	11.59	**2.62**	147.46	**6.26**	11.27	-	-	**0.28**
V2A	AudioX [16]	3.13	18.90	4.01	205.48	**5.87**	8.31	1.20	0.23	0.13
	MMAudio [15]	2.94	13.41	3.87	159.30	5.50	7.59	0.62	0.24	0.14
	Omni2Sound (ours)	**2.47**	**8.78**	**2.55**	**112.21**	5.78	**8.56**	**0.57**	**0.34**	**0.18**
VT2A	ThinkSound ($w/o.$ CoT) [14]	2.53	11.99	3.52	206.93	5.77	6.09	0.66	0.22	0.19
	HunyuanVideo-Foley [13]	2.13	8.06	3.58	**94.64**	**6.04**	8.17	**0.55**	**0.34**	0.23
	AudioX [16]	2.39	14.26	3.16	149.37	5.97	**10.23**	1.21	0.23	0.26
	MMAudio [15]	2.41	10.12	4.90	129.21	5.53	7.46	0.59	0.25	0.20
	Omni2Sound (ours)	**2.10**	**7.60**	**2.37**	106.55	5.98	8.22	0.58	0.32	**0.26**

dioX [16], MMAudio [15]) and specialized models (e.g. ThinkSound [14], HunyuanVideo-Foley [13]). To further validate Omni2Sound's generalization beyond our SoundAtlas captioning style, we evaluate it on the same VGGSound test clips but use the Video-LLaMA [40] captions from ThinkSound [14]. As shown in Table 3 (w/ Video-LLaMA caps), while performance slightly degrades, our model's scores still surpass all baselines, confirming its robustness to unseen captioning styles.

Generalization on Third-Party Benchmarks. To validate generalization, we evaluate on Kling-Audio-Eval [30] and AudioCaps [24] results in Table 4 and Appendix Table 7. On Kling-Audio-Eval, Omni2Sound remains highly competitive despite the domain gap (YouTube-sourced SoundAtlas vs. Kling's professional video). While trailing HunyuanVideo-Foley [13] in some metrics, which is expected given its massive data advantage (100k vs 2k hours), our model consistently outperforms other unified and specialized baselines across all tasks. Furthermore, on AudioCaps, Omni2Sound achieves top-tier performance against specialized T2A models, securing the best scores in distribution metrics (KL, FD) and semantic alignment (CLAP $= 0.36$), while remaining highly competitive in audio quality (PQ) and the FAD metric.

Subjective Evaluation. To validate perceptual performance, we conduct a human evaluation (detailed in Appendix F) across three dimensions: Acoustic Fidelity (MOS-Q), Semantic Consistency (MOS-S), and Temporal Synchronization (MOS-T). As shown in Appendix Fig. 4, Omni2Sound outperforms all baselines on both VT2A and V2A tasks. Crucially, these subjective results are highly consistent with the objective metrics in Table 3, confirming our model's superiority in both generation quality and cross-modal alignment.

6.3. Ablation Studies

We first analyze the multi-task training dynamics in Table 5 to demonstrate how high-quality data resolves task competition, and then use Table 6 to prove the necessity of our three-stage progressive training schedule.

High-Quality VT2A Data as a Critical Bridge. We first investigate the Cross-Task Competition between V2A and T2A, which persists even when models are based on the T2A pretraining from Stage 1. As shown in Table 5 (rows 1-2), a naive joint training of V2A and T2A results in a severe trade-off. Increasing the T2A sampling ratio (π_{T2A}) from 0.20 to 0.40 improves T2A performance (FAD 1.36 → 1.06) but simultaneously degrades V2A generation (FAD 0.56 → 0.62), preventing simultaneous optimization. Our insight is that this conflict is resolved by introducing high-quality VT2A data as a critical bridge. This hypothesis is validated in row 3, which introduces our SoundAtlas data (denoted by TA* and VTA*). The results show a dramatic performance boost, achieving the best metrics across all tasks (e.g., T2A FAD 0.94, V2A FD 3.61, VT2A FD 2.83). This confirms that the high V-A-T alignment in SoundAtlas is essential for resolving the V2A-T2A competition and fostering a cooperative dynamic.

To further emphasize that this bridging effect is contingent on data quality, we provide a comparison in row 4. Here, we use standard-quality data (TA/VTA), where captions were generated by Gemini-2.5 using only the audio modality. Although the VT2A task is present, the poor V-T-A alignment fails to resolve the competition, and performance is still severely compromised (e.g., T2A FAD 1.13), far underperforming the SoundAtlas-driven model. This comparison proves that it is not merely the VT2A task, but the high-fidelity alignment of the bridge data, that is essential. This high quality enables data efficiency: the T2A ratio can be dropped to $\pi_{T2A} = 0.1$ while achieving SOTA T2A performance, mitigating resource contention as designed.

Table 5. Ablation study on the Stage 2 multi-task training strategy. TA*/VTA* denotes data from our high-alignment SoundAtlas dataset, while TA/VTA denotes data from a baseline with audio-only captions generated by Gemini 2.5.

Training Strategy	$\pi_{T2A} : \pi_{V2A} : \pi_{VTA}$	T2A Task		V2A Task				VT2A Task			
		FAD↓	FD↓	FAD↓	FD↓	DS↓	IB↑	FAD↓	FD↓	DS↓	IB↑
TA+VA	0.20 : 0.80 : 0.00	1.36	5.52	0.56	4.13	0.50	0.33	-	-	-	-
TA+VA	0.40 : 0.60 : 0.00	1.06	4.62	0.62	4.63	0.52	0.32	-	-	-	-
TA*+VA+VTA*	0.10 : 0.35 : 0.55	0.94	4.22	0.57	3.61	0.49	0.33	0.53	2.83	0.51	0.32
TA+VA+VTA	0.20 : 0.30 : 0.50	1.13	4.68	0.56	4.22	0.50	0.32	0.62	3.51	0.51	0.33

Table 6. Ablation study on our progressive multi-task training. We compare our full S1 → S2 → S3 model against three baselines (S2, S1 → S2, and S1 → [S2+S3]). All models are trained for the same total 1.2M steps.

Task	Method	FAD↓	FD↓	DS↓	IB↑
T2A	S2	1.22	5.88	-	-
	S1 → S2	0.94	4.62	-	-
	S1 → [S2+S3]	1.11	4.45	-	-
	S1 → S2 → S3	1.01	4.61	-	-
V2A	S2	0.68	4.70	0.47	0.33
	S1 → S2	0.57	3.61	0.49	0.33
	S1 → [S2+S3]	0.60	3.81	0.47	0.34
	S1 → S2 → S3	0.51	3.41	0.47	0.35
VT2A	S2	0.63	4.40	0.49	0.33
	S1 → S2	0.53	2.83	0.51	0.32
	S1 → [S2+S3]	0.61	3.27	0.50	0.33
	S1 → S2 → S3	0.53	2.95	0.49	0.34

Necessity of the Progressive Three-Stage Schedule. Next, we demonstrate the necessity of our full progressive schedule in Table 6. We compare our full S1 → S2 → S3 pipeline against three baselines, all trained for the same total steps on SoundAtlas data. First, comparing the S2 only model with the S1 → S2 model confirms the value of the Stage 1 generative prior. Without S1, the S2 only model fails to converge well, showing poor quality (T2A FAD 1.22, V2A FAD 0.68). The S1 → S2 model, benefiting from the pretraining, significantly boosts generation quality (T2A FAD 0.94, V2A FAD 0.57) and resolves the Cross-Task Competition. However, this model still suffers from Intra-Task Competition (modality bias), as evidenced by its weaker A-V synchronization (V2A DS 0.49). Second, we validate our crucial hypothesis that Stage 3 must be decoupled. The S1 → [S2+S3] baseline, which merges the S3 robustness augmentations directly into S2, destabilizes the fragile optimization process. While it maintains A-V synchronization (V2A DS 0.47), introducing these augmentations prematurely harms the generative quality achieved in S2, leading to a clear degradation in FAD/FD scores (e.g., V2A FAD 0.60, VT2A FAD 0.61).

Finally, our full S1 → S2 → S3 model resolves both challenges. As established in our method, S3 has two complementary goals: mitigating the text bias (via Text Dropout) and the video bias (via Off-screen Synthesis).

Table 7. VT2A evaluation on VGGSound-Omni off-screen track. We compare S1→S2 against our full S1→S2→S3 model to validate *Off-screen Synthesis* augmentation.

Method	FAD↓	KL↓	LA-CLAP↑	Win Rate↑
S1 → S2	0.97	1.46	0.31	46.8%
S1 → S2 → S3	0.85	1.39	0.32	53.2%

The main results in Table 6 confirm the first goal: the full S3 model enhances cross-modal consistency (V2A DS 0.49 → 0.47) while achieving the highest overall generation quality (V2A FAD 0.51). To validate the second goal—improving faithfulness against a video bias—we conduct a targeted evaluation on our VGGSound-Omni off-screen track, presented in Table 7. This table compares the S1→S2 baseline against our full model, showing the S3 augmentations yield superior audio quality and improved objective text-audio alignment. This gain in faithfulness is further confirmed by a subjective preference test using an MLLM-as-Judge (evaluating text-audio faithfulness on a 1-to-5 scale).

7. Conclusion

In this work, we addressed the foundational challenges of unified video-text-to-audio (VT2A) generation: data scarcity and cross-task competition. We introduce a three-part contribution: SoundAtlas, the first large-scale, human-expert-level audio caption dataset; Omni2Sound, a unified model featuring a three-stage progressive schedule to resolve task competition; and VGGSound-Omni, a comprehensive benchmark for unified VT2A evaluation. Our experiments demonstrate that this approach effectively resolves cross-task and intra-task competition and enables Omni2Sound to achieve unified state-of-the-art performance.

Acknowledgments

This work is supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM101), the National Natural Science Foundation of China (62550004, U24A20342, U25B6003, 92570001), and the Australian Research Council (DP260100218).

References

[1] F. Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre D'efossez, et al. Audiogen: Textually guided audio generation. *ArXiv*, abs/2209.15352, 2022. 1

[2] 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. *arXiv preprint arXiv:2301.12503*, 2023. 6, 4

[3] Zach Evans, Julian Parker, CJ Carr, Zack Zukowski, Josiah Taylor, et al. Stable audio open. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2024. 5, 3, 4

[4] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction-tuned llm and latent diffusion model. *ArXiv*, abs/2304.13731, 2023.

[5] Yuxuan Jiang, Zehua Chen, Zeqian Ju, Yusheng Dai, Weibei Dou, and Jun Zhu. Controlaudio: Tackling text-guided, timing-indicated and intelligible audio generation via progressive diffusion modeling. *arXiv preprint arXiv:2510.08878*, 2025.

[6] Yuxuan Jiang, Zehua Chen, Zeqian Ju, Chang Li, Weibei Dou, and Jun Zhu. Freeaudio: Training-free timing planning for controllable long-form text-to-audio generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025. 1

[7] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *ArXiv*, abs/2306.17203, 2023. 1

[8] Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation network with rectified flow matching. *Advances in neural information processing systems*, 37:128118–128138, 2024. 6

[9] Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, et al. Foleycrafter: Bring silent videos to life with lifelike and synchronized sounds. *International Journal of Computer Vision*, 134, 2024.

[10] Ziyang Chen, Prem Seetharaman, Bryan Russell, Oriol Nieto, David Bourgin, et al. Video-guided foley sound generation with multimodal controls. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18770–18781, 2024. 1

[11] Saksham Singh Kushwaha and Yapeng Tian. Vintage: Joint video and text conditioning for holistic audio generation. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13529–13539, 2024. 1, 2

[12] Rongjie Huang, Dongchao Yang, Huadai Liu, Xixin Wu, and Helen M. Meng. Reasonaudio: Semantic reasoning and temporal synchrony in video–text-to-audio generation, 2025. 1

[13] Sizhe Shan, Qiulin Li, Yutao Cui, Miles Yang, Yuehai Wang, et al. Hunyuanvideo-foley: Multimodal diffusion with representation alignment for high-fidelity foley audio generation. *ArXiv*, abs/2508.16930, 2025. 1, 2, 6, 7, 3

[14] Huadai Liu, Jialei Wang, Kaicheng Luo, Wen Wang, Qian Chen, et al. Thinksound: Chain-of-thought reasoning in multimodal large language models for audio generation and editing. *ArXiv*, abs/2506.21448, 2025. 1, 6, 7

[15] Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander G. Schwing, et al. Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28901–28911, 2024. 1, 2, 3, 5, 6, 7, 4

[16] Zeyue Tian, Yizhu Jin, Zhaoyang Liu, Ruibin Yuan, Xu Tan, et al. Audiox: Diffusion transformer for anything-to-audio generation. *ArXiv*, abs/2503.10522, 2025. 1, 2, 3, 5, 6, 7

[17] Liyang Chen, Hongkai Chen, Yujun Cai, Sifan Li, Qingwen Ye, et al. Detecting and mitigating insertion hallucination in video-to-audio generation. *ArXiv*, abs/2510.08078, 2025. 2

[18] Honglie Chen, Weidi Xie, A. Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725, 2020. 2, 3, 4, 5, 6

[19] J. Gemmeke, D. Ellis, Dylan Freedman, A. Jansen, W. Lawrence, et al. Audio set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. 2, 3, 4, 6

[20] Gemini Team. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805, 2023. 2, 3, 4, 1

[21] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 4

[22] Daniil Zverev, Thaddaus Wiedemer, Ameya Prabhu, Matthias Bethge, Wieland Brendel, et al. Vggsounder: Audio-visual evaluations for foundation models. *ArXiv*, abs/2508.08237, 2025. 2, 5, 6

[23] William S. Peebles and Saining Xie. Scalable diffusion models with transformers. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182, 2022. 2, 4, 5

[24] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. pages 119–132, 2019. 2, 4, 6, 7

[25] K. Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: an audio captioning dataset. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740, 2019. 2, 6, 4

[26] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, et al. 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, 2023. 2, 6, 4

[27] Jisheng Bai, Haohe Liu, Mou Wang, Dongyuan Shi, Wenwu Wang, et al. Audiosetcaps: An enriched audio-caption dataset using automated generation pipeline with large audio and language models. *IEEE Transactions on Audio, Speech and Language Processing*, 33:2817–2829, 2024. 2, 3, 4

[28] Luoyi Sun, Xuenan Xu, Mengyue Wu, and Weidi Xie. Auto-acd: A large-scale dataset for audio-language representation learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5025–5034, 2024. 3, 4

[29] Yi Yuan, Dongya Jia, Xiaobin Zhuang, Yuanzhe Chen, Zhuo Chen, Yuping Wang, Yuxuan Wang, Xubo Liu, Xiyuan Kang, Mark D Plumbley, et al. Sound-vecaps: Improving audio generation with visually enhanced captions. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025. 3, 4

[30] Le Wang, Jun Wang, Chunyu Qiang, Feng Deng, Chen Zhang, et al. Audiogen-omni: A unified multimodal diffusion transformer for video-synchronized audio, speech, and song generation. *ArXiv*, abs/2508.00733, 2025. 3, 6, 7, 2

[31] Xuenan Xu, Jiahao Mei, Zihao Zheng, Ye Tao, Zeyu Xie, et al. Uniflow-audio: Unified flow matching for audio generation from omni-modalities. *ArXiv*, abs/2509.24391, 2025. 3

[32] Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*, 2025. 3

[33] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, et al. Qwen3-omni technical report. *CoRR*, abs/2509.17765, 2025. 3

[34] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. 4, 6

[35] Zach Evans, CJ Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. *ArXiv*, abs/2402.04825, 2024. 4

[36] Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, et al. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416, 2022. 4

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, et al. Learning transferable visual models from natural language supervision. pages 8748–8763, 2021. 4

[38] Vladimir E. Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5325–5329, 2024. 5

[39] Yusheng Dai, Chenxi Wang, Chang Li, Chen Wang, Jun Du, Kewei Li, Ruoyu Wang, Jiefeng Ma, Lei Sun, and Jianqing Gao. Latent swap joint diffusion for 2d long-form latent generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11006–11015, 2025. 5

[40] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. pages 543–553, 2023. 6, 7

[41] Ilpo Viertola, Vladimir E. Iashin, and Esa Rahtu. Temporally aligned audio for video with autoregression. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2024. 6

[42] OpenAI. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/, 2025. 6

[43] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, et al. Imagebind one embedding space to bind them all. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, 2023. 6, 4

[44] A. Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, et al. Musiclm: Generating music from text. *ArXiv*, abs/2301.11325, 2023. 6, 2

[45] Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. FSD50K: an open dataset of human-labeled sound events. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30:829–852, 2022. 6, 4

[46] Thierry Bertin-Mahieux, Daniel PW Ellis, Brian Whitman, and Paul Lamere. The million song dataset. 2011. 6, 4

[47] Michaël Defferrard, Kirell Benzi, P. Vandergheynst, and X. Bresson. Fma: A dataset for music analysis. pages 316–323, 2016. 6, 4

[48] Chenxi Wang, Yusheng Dai, Lei Sun, Jun Du, and Jianqing Gao. Audioatlas: A comprehensive and balanced benchmark towards movie-oriented text-to-audio generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, 2025. 6

[49] Shawn Hershey, Sourish Chaudhuri, D. Ellis, J. Gemmeke, A. Jansen, et al. Cnn architectures for large-scale audio classification. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2016. 6, 4

[50] Khaled Koutini, Jan Schlüter, Hamid Eghbalzadeh, and G. Widmer. Efficient training of audio transformers with patchout. *ArXiv*, abs/2110.05069, 2021. 6, 4

[51] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, et al. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2019. 6, 4

[52] Tim Salimans, I. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, et al. Improved techniques for training gans. *ArXiv*, abs/1606.03498, 2016. 6, 4

[53] Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, et al. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *ArXiv*, abs/2502.05139, 2025. 6

[54] Yusong Wu, K. Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, et al. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2022. 6, 4

[55] Vladimir E. Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5325–5329, 2024. 6, 4

[56] 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. 2

[57] Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, et al. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024. 2

[58] Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao. Make-an-audio 2: Temporal-enhanced text-to-audio generation. *arXiv preprint arXiv:2305.18474*, 2023. 2

[59] Moayed Haji-Ali, Willi Menapace, Aliaksandr Siarohin, Guha Balakrishnan, and Vicente Ordonez. Taming data and transformers for audio generation. *International Journal of Computer Vision*, 134(3):87, 2026. 2