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

Supplementary Material

Overview This document provides technical details, evaluation protocols, and extended experimental analyses. We begin with the **Cost Analysis** in Section A, validating SoundAtlas as a scalable and cost-effective pipeline. We then provide the exact **Audio Caption Prompt Instructions** in Section B, followed by detailed **Evaluation Protocols** to compare the quality of Audio Caption Datasets in Section C and the detailed construction of the **Off-Screen Benchmark Track** in Section D. Furthermore, we demonstrate the model's **Generalization Capabilities on third-party benchmarks** in Section E and elaborate on the **User Study** in Section F. Section G outlines the **Implementation Details**, including model configurations and training data composition. Section H defines the **Objective Evaluation Metrics on Generation Audio** used throughout the paper. **Qualitative results** can be found in static HTML file.

A. Cost Analysis on Audio Captioning

While Gemini 2.5 Pro [20] represents a milestone as a native multimodal foundation model, utilizing it directly for large-scale video-grounded audio captioning proves economically unsustainable. As quantified in Table 6, using Gemini's standard API pricing, a naive implementation—processing raw video frames alongside audio ($V + A$)—incurs a prohibitive expenditure of $10,275 USD per 1M samples. This figure is derived from the token consumption of a 10-second sample: the input aggregates to 3,820 tokens (comprising 1,000 instruction, 320 audio, and 2,500 visual tokens), while the full chain-of-thought generation requires $\sim$550 output tokens. Crucially, this naive approach suffers from an inherent visual bias, as shown in Figure 1 in main paper.

To address these challenges, our SoundAtlas pipeline employs three strategic optimizations. First, we implement *Vision-to-Language Compression*. This strategy replaces expensive raw video with a concise video caption c_v, eliminating the large $\sim$2,500 token visual overhead (Table 6, Row 2) and effectively mitigating the visual modality bias. Second, we enforce *Restricted Reasoning*, capping the generation output at $\sim$160 tokens (Table 6, Row 3). Finally, we utilize a *Junior-Senior Agent Handoff* that defaults to the cost-effective Flash model G_{junior} for the majority of samples, reserving the Senior agent (G_{senior}) solely for complex cases. As shown in Table 6, while the standalone Flash model offers the lowest theoretical cost ($1,026), our hybrid pipeline strikes a balance between quality and efficiency, reducing the initial expenditure of $10,275 to approximately $2,000 per million samples.

B. Audio Caption Prompt Instructions

As illustrated in Figure 5, we present the audio captioning system prompt employed in our agentic annotation pipeline to construct the *SoundAtlas* dataset.

C. Audio Caption Dataset Comparison

We provide the detailed scoring process for both MLLM-as-a-judge and Human Expert Evaluation on different audio caption datasets in Table 1 and 2 of main paper. The evaluation methodology consists of two stages: (1) absolute scoring based on the specific linguistic criteria defined below, and (2) a comparative win-rate calculation derived from these scores.

Subjective Evaluation Protocol. We formulate a standardized scoring protocol for both MLLM and human evaluators, focusing on two distinct dimensions of modality alignment.

1. Semantic Alignment (MOS-S, Scale 1-4). This metric assesses both *Accuracy* (factuality of sound events) and *Detail* (precision of adjectives). The scale is defined as: (1) Factually incorrect/Brief; (2) Mostly incorrect/Brief; (3) Minor errors/Detailed (but visually redundant); and (4) Error-free and Detailed (strictly audio-centric).

2. Temporal Alignment (MOS-T, Scale 1-3). This evaluates whether the chronological order of described events matches the audio stream. The scale ranges from (1) Disordered, (2) Partially Correct, to (3) Perfectly Ordered. Samples with constant or stationary sounds (lacking distinct temporal events) are marked as `N/A` and excluded from this metric.

Human Evaluation Setup. To complement and validate our automated evaluation, we conducted a dedicated human expert evaluation based on the aforementioned protocol. We randomly sampled a subset of 100 instances from the evaluation corpus used in the MLLM-as-a-judge benchmark. We recruited five expert annotators with professional backgrounds in audio-visual analysis to assess these samples independently. To ensure robustness and mitigate individual bias, the final score for each item is derived by calculating the average rating across the five evaluators. For reference, the user study interface is illustrated in Figure 6.

Win Rate Calculation. We adopt a general pairwise comparison paradigm. For each evaluation set, a target

Table 6. Cost Analysis on Audio Captioning with Gemini 2.5. We compare the inference costs for processing one million 10-second samples. The table demonstrates a step-by-step ablation path: removing raw video (Row 2), restricting reasoning with vision-to-language compression (Row 3), and switching to the Flash model (Row 4) progressively reduces costs from $10,275 to $1,026.

Model Configuration	Input Modality	Input Token Num.	Output Token Num.	Est. Cost (USD / 1M Samples)
Gemini 2.5 Pro (Thinking-Full)	T + V + A	3,820	550	$10,275.00
Gemini 2.5 Pro (Thinking-Full)	T + A	1,340	550	$7,175.00
Gemini 2.5 Pro (Thinking-128)	T + A	1,340	160	$3,275.00
Gemini 2.5 Flash (Thinking-128)	T + A	1,340	160	$1,026.00

Table 7. Comparison of the generation performance on unified VT2A models and T2A models on Audiocaps test set.

Method	KL↓	FD↓	FAD↓	PQ↑	LA-CLAP↑
AudioLDM 2-L [56]	1.73	34.21	2.26	**5.93**	0.24
TANGO 2 [57]	**1.19**	15.92	3.17	5.82	<u>0.35</u>
Make-An-Audio 2 [58]	1.38	15.34	<u>1.46</u>	5.64	0.25
GenAU-Large [59]	1.42	16.92	**1.32**	5.52	0.26
MMAudio [15]	1.43	<u>13.78</u>	2.92	5.30	0.29
AudioX [16]	1.55	17.10	2.65	5.81	0.31
Omni2Sound (Ours)	<u>1.35</u>	**11.42**	1.74	<u>5.84</u>	**0.36**

model is compared against an opposing method. The Mean Win Rate (MWR) for any given model is derived by aggregating the outcomes of all its pairwise comparisons:

$$\text{MWR} = \frac{N_{\text{win}} + 0.5 \times N_{\text{tie}}}{N_{\text{total}}} \quad (1)$$

where N_{win}, N_{tie}, and N_{total} denote the number of wins (scoring 1.0), ties (scoring 0.5), and total pairwise comparisons involving that model, respectively.

D. Off-Screen Track of VGGSound-Omni

We introduce a dedicated Off-Screen Audio-Generation Track of VGGSound-Omni. This subset specifically evaluates the model's capacity to handle non-depicted audio sources and is constructed through two distinct pipelines: (i) a *Natural Off-screen Events* subset sourced from the original test set; and (ii) a *Synthetic Music* subset focusing on background music (BGM) generation.

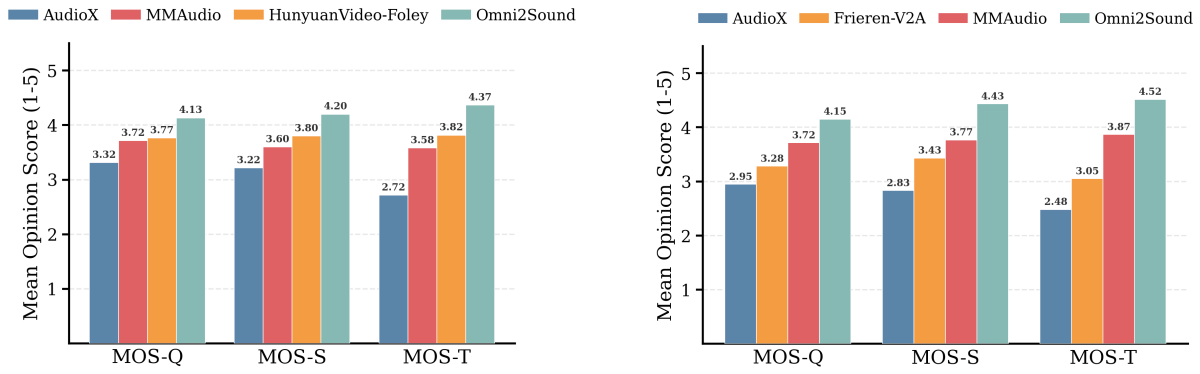
Natural Off-screen Events. We construct the *Natural Events* subset by identifying VGGSound clips that inherently contain off-screen audio cues. The curation involves a rigorous three-step filtering pipeline. First, regarding Metadata & Modality, we ensure acoustic purity by excluding samples with pre-existing background music, static imagery, or voice-overs. Crucially, we filter out videos containing vision-only ("V") labels, retaining only those with Audio-Visual ("AV") or Audio-only ("A") modalities. Second, for Complexity & Consistency, we limit scene complexity to a maximum of 6 labels. To capture "natural" off-screen scenarios, we filter based on the AV Ratio—defined as the proportion of "AV" labels relative to the total label count. We explicitly select samples where this ratio falls within $[0.25, 0.80]$, ensuring that the audio content is not perfectly aligned with the visual stream (i.e., low A-V correspondence). Finally, we apply Distribution Balancing to mitigate the over-representation of common classes, restricting the proportion of speech to 20%.

Synthetic Music Augmentation. To address the high demand for Background Music (BGM) generation, we create a *Synthetic Music* subset by mixing semantically aligned MusicCaps [44] clips into a pool of high-fidelity videos. This process follows a two-stage procedure. In the Base Selection stage, we first select a "clean" video pool by strictly requiring a 100% AV label ratio and filtering for high alignment (ImageBind ≥ 0.30, Desync < 0.55), ensuring all original acoustic events are visually manifest. Subsequently, during Semantic Mixing, we augment these videos with background music tracks. To guarantee semantic coherence, we utilize GPT to retrieve the most congruent music track from a random candidate batch of 50 samples based on the video context. Ground-truth captions are updated to reflect this acoustic addition.

Comparison with Concurrent Work. We acknowledge the pioneering work of VinTAGe-Bench [11] in synthetic robustness evaluation. However, the off-screen subset of our VGGSound-Omni benchmark extends this direction in three critical dimensions. First, in terms of **Realism**, by leveraging VGGSounder [22] metadata, our natural subset is primarily sourced from real-world off-screen audio events rather than relying solely on synthetic mixes. Second, regarding **Scale**, our benchmark is significantly larger, providing 1,613 evaluation items compared to the 212 basic videos of VinTAGe-Bench. Third, regarding **Scope**, we include a dedicated *Synthetic Music (BGM)* track, addressing a critical, high-demand scenario often overlooked in standard environmental sound benchmarks.

E. Generalization on Third-Party Benchmarks.

To further validate our model's generalization and mitigate potential biases from our self-constructed benchmark, we evaluate it on the Kling-Audio-Eval [30] and Audiocaps test set [24]. In Table 4, on the Kling-Audio-Eval benchmark, Omni2Sound remains highly



(a) Subjective Evaluation Results on VT2A Task

(b) Subjective Evaluation Results on V2A Task

Figure 4. Subjective Evaluation Results on VGGSound-Omni. We report Mean Opinion Scores (MOS) on a 1-5 scale across three dimensions: Acoustic Quality (MOS-Q), Semantic Alignment (MOS-S), and Temporal Alignment (MOS-T). Omni2Sound consistently outperforms competitive baselines (AudioX, MMAudio, HunyuanVideo-Foley, Frieren-V2A) across all perceptual metrics on both VT2A and V2A tasks, validating its superior generation fidelity and alignment.

competitive, despite a significant data scale and distribution gap (our YouTube-sourced SoundAtlas vs. Kling's professional video/Foley). While HunyuanVideo-Foley [13] leads on several metrics, this is expected given its massive 100k-hour internal dataset, which is tens of times larger than our SoundAtlas filter derived from VGGSound and AudioSet. Nevertheless, Omni2Sound consistently outperforms all other strong baselines (e.g., MMAudio, AudioX, and ThinkSound) across V2A and VT2A tasks, demonstrating strong generalization as the SOTA or second-best method. In Table 7, on the Audiocaps test set, we compare Omni2Sound against specialized SOTA T2A models. The results show our unified model achieves top-tier performance, attaining the best scores in key distribution metrics (KL, FD) and semantic alignment (CLAP $= 0.36$), while remaining highly competitive in audio quality (PQ) and the FAD metric.

F. User Study

We conduct a comprehensive user study on the VGGSound-Omni benchmark to validate Omni2Sound against top baselines (four methods in total). Given the density of comparisons involved, we structure VT2A and V2A as independent evaluation tracks to mitigate evaluator fatigue. We recruit a total of 16 expert evaluators, who are evenly distributed across the two independent tasks. Each participant evaluates 20 random samples (80 comparisons) within their assigned track. Samples from the same source are grouped with randomized method order to maintain blinding. In total, 1280 responses per metric are collected.

Subjective Evaluation Metrics. Our final evaluation utilizes a multi-dimensional Mean Opinion Score (MOS) protocol, where expert human evaluators assess the generated audio across three distinct criteria. All scores are normalized to a 5-point Likert scale (1: Poor/Misaligned; 5: Excellent/Perfectly Aligned).

- **MOS-Q: Acoustic Fidelity (Quality).** This metric assesses the intrinsic acoustic quality and perceptual realism of the generated sound, independent of the conditioning inputs. Evaluators focus on auditory naturalness, clarity, and the absence of technical artifacts (e.g., distortion, noise, mixing comfort).
- **MOS-S: Semantic Consistency (Alignment).** This quantifies the perceptual fidelity between the content of the generated audio and the semantic information conveyed by the conditioning modalities (video frames and textual captions). Evaluation centers on whether the generated sound event's category and characteristics logically correspond to the depicted visual and textual context.
- **MOS-T: Temporal Synchronization (Alignment).** This assesses the temporal accuracy of the acoustic events against the visual stream. Evaluators specifically check the precision of sound onset, offset, and duration, ensuring tight synchronization with the corresponding visual event timing.

The results, summarized in Figure 4, demonstrate that Omni2Sound outperforms all baselines across the three subjective metrics: MOS-Q, MOS-S, and MOS-T on both VT2A and V2A tasks. This strong alignment between human preference in Figure 4 and the objective metrics presented in Table 3 in main paper validates the effectiveness of our proposed data construction and training pipeline. For reference, the user study interface is illustrated in Figure 7.

G. Implementation Details

Model Configuration. Following Stable Audio [3], our diffusion model adopts **a** Diffusion Transformer (DiT) architecture within a Latent Diffusion Model

(LDM) paradigm. The diffusion backbone consists of a DiT with 24 layers, 24 attention heads, and a hidden dimension of 1536. We employ cross-attention mechanisms to inject semantic conditions (e.g., FLAN-T5 and CLIP embeddings) and Adaptive Layer Normalization (AdaLN) to integrate temporal signals, as detailed in Section 4.1. Both the conditional token dimension and the global condition embedding dimension are 1024. Finally, for audio compression, we train a Variational Autoencoder (VAE) from scratch based on the wav Audio VAE architecture [3], operating at a 16kHz sampling rate. With strides of $[4, 4, 4, 10]$, the encoder achieves a total downsampling ratio of 640, mapping mono waveforms into a compact 64-dimensional latent space. To ensure high-fidelity reconstruction, we utilize Snake activations throughout the network.

Training Data. 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]. All audio signals are standardized to a mono-channel format at 16kHz. To accommodate fixed-size diffusion inputs, we normalize clips to a uniform 10-second duration: samples exceeding this length undergo right cropping, while shorter samples are right-padded with silence.

Subsequently, the model is fine-tuned for unified multimodal tasks using our proposed SoundAtlas. Constructed following the pipeline **detailed in** Section 5, this dataset comprises 470k high-quality V-A-T pairs, sourced from 140k VGGSound and 330k AudioSet samples. Notably, the AudioSet subset is strictly curated: starting from the original 2M corpus, we first applied a preliminary filtration to exclude all speech- and music-related categories, resulting in a candidate pool of 450k sound samples. These candidates then underwent our A-V consistency routing and verification pipeline to yield the final 330k high-fidelity pairs. For T2A task fine-tuning, we augment the training with T-A pairs from SoundAtlas as well as a high-fidelity subset of the pre-training corpus, filtered by strict quality thresholds: requiring a CLAP score greater than 0.35 and a PQ score exceeding 6.0.

H. Objective Evaluation Metrics.

We implement our objective evaluation metrics using the standardized AV-benchmark toolkit [15]. All samples are generated under the same video and text conditions and evaluated in 8-second clips, following previous work [15]. Following common practice [2], we assess the quality of the generation in four critical dimensions. For Distribution Matching, we measure the similarity in feature distribution between generated and ground-truth audio. We compute the Fréchet Distance using the VGGish (FAD) [49] and PaSST ($\mathrm{FD}_{\mathrm{PaSST}}$) [50] embeddings, as well as the Fréchet Audio Distance using PANNs (FD) [51]. We also report the Kullback-Leibler divergence using PANNs (KL) and PaSST ($\mathrm{KL}_{\mathrm{PaSST}}$) classifiers. For Audio Quality, we assess the quality of the generation using the Inception Score [52], calculated with both the PANNs (IS) and PaSST ($\mathrm{IS}_{\mathrm{PaSST}}$) classifiers. For Semantic Alignment, we evaluate text-audio consistency using LAION CLAP (CLAP) [34] and Microsoft CLAP (MS-CLAP) [54] scores, and video-audio alignment using ImageBind score (IB) [43] as cosine similarity between video and audio embeddings. Finally, for Temporal Alignment, we assess audio-visual synchrony using the DS metric predicted by Synchformer [55].

Audio Captioning Instruction for SoundAtlas

Roles and Tasks

You are an experienced audio content analyst skilled in describing soundscapes through detailed, multi-dimensional natural language. Given an audio clip (*a*) and its corresponding video descriptions (T_v), identify and describe all relevant auditory elements in chronological order, then write a rich audio description that faithfully and dynamically reflects the scene.

Annotation Dimensions

1. **Primary Sound Information**
   - **Humans/Animals:** speech (talking, shouting), movements (footsteps). *Note: Do not transcribe words/lyrics; describe voice characteristics.*
   - **Objects:** traffic, office sounds, battlefield, tools.
   - **Characteristics:** Gender/age, language, quantity (monologue/turn-taking), emotional tone, voice qualities.
2. **Background Sounds (if present)**
   - Natural (wind, rain) or Artificial (city noise, crowds). Briefly specify the environment if necessary.
3. **Music (if present)**
   - Style/genre, rhythmic features, emotional tone, atmosphere.
   - Identifiable instruments and effects (harmonies, reverb).
4. **Detailed Descriptors**
   - Changes in volume/speed/intensity. Narrative functions.
   - Detailed duration, spatial distance, pitch, timbre, texture.

Important Guidelines

1. **Avoid Redundancy:** Identify sources once unless they change significantly. Keep it concise.
2. **Prioritize the Audio:** Use video description *only* to clarify ambiguous sounds. If a sound isn't audible, don't describe it.
3. **Avoid Hallucinated Sounds:** Only describe perceptible sounds. Avoid describing artifacts (e.g., "high-pitched squeal" from edits).

Output Format

Integrate elements into **one or few sentences** following these rules:

- **Language:** English.
- **Structure:** No lists or bullet points.
- **Length:** Max 40 words. Concise but detailed.
- **Temporal Order:** Chronological (e.g., "first", "then", "suddenly").
- **Style:** Natural, objective, context-sensitive. Focus on what is heard.

Examples

Example 1 (General):

> **Input:** [High-pitched mechanical whirring with periodic thuds]
> **Video Caption:** "Laundromat with washing machines and dryers running"
> **Output:** Washing machines whir at high speed while dryers tumble clothes with periodic rhythmic thuds. Water drains intermittently as cycles complete and doors slam shut.

Example 2 (Anti-hallucination):

> **Input:** [Guitar strumming and melody]
> **Video Caption:** "Musician performing with piano and guitar on stage"
> **Output:** Acoustic guitar plays melodic fingerpicking patterns with clear, resonant tones. *(Piano is omitted as it is not audible).*

Figure 5. Audio Captioning Instruction for SoundAtlas.

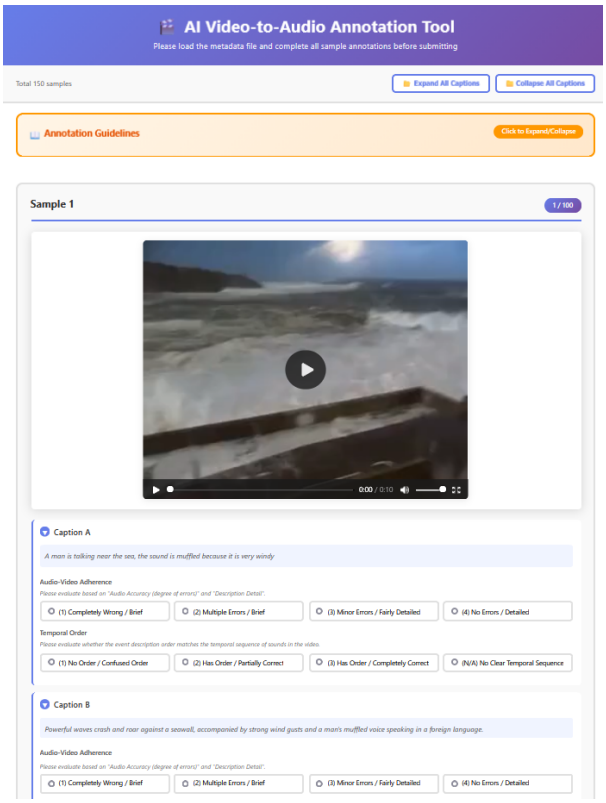


Figure 6. User study interface for human evaluation across different audio generation models.

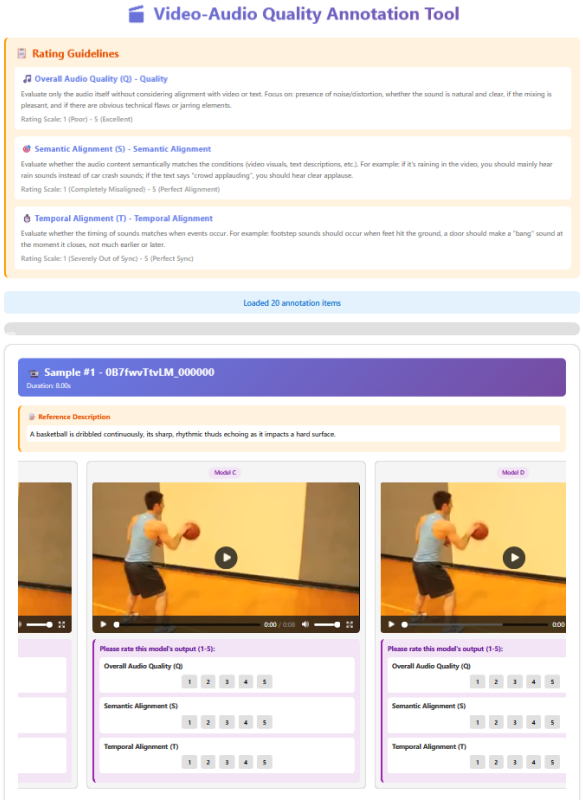


Figure 7. User study interface for human evaluation across different automatic audio captioning datasets.