

# LLaSO - Supplementary Material (Anonymous Submission)

---

## 1. Anonymized repository

For double-blind review, we provide an anonymized code repository:

<https://anonymous.4open.science/r/LLaSO-DEBB/>

This repo removes or neutralizes any links and metadata that could reveal author identity. It contains minimal scripts and instructions sufficient to reproduce the evaluation pipeline on the provided subset without leaking identities.

---

## 2. Why only `LLaSO-Eval-mini.zip` is provided

To keep the submission anonymous and lightweight, we include a **minimal working subset** of our benchmark:

- The subset preserves the **same data schema, modality configurations, and task types** as the full benchmark.
  - It is sufficient to **run the evaluation/inference pipeline end-to-end** and validate data loading, I/O format, and code paths.
- 

## 3. Relationship between Eval / Align / Instruct

Our evaluation benchmark (**LLaSO-Eval**) is designed to **mirror** the modality configurations and task coverage used in training data:

- **Modality configurations:**
  - `text_audio/` – text instruction + audio input
  - `audio_text/` – audio instruction + text input
  - `pure_audio/` – audio-only

- **Task taxonomy** matches training (LLaSO-Instruct) and the alignment data (LLaSO-Align), enabling consistent training-evaluation mapping.

Therefore, providing a **minimal subset of LLaSO-Eval** is enough for reviewers to verify format compatibility and pipeline functionality without shipping the full training sets during double-blind review.

---

## 4. What's inside **LLaSO-Eval-mini.zip**

After extraction, you will get a directory like:

```
1  LLaSO-Eval-mini/  
2    └─ audio_text/ *.json (task name; 2 samples each json)  
3    └─ pure_audio/  
4    └─ text_audio/  
5    └─ source data name/ *.wav (2 samples)
```

---

## 5. How to use the subset

You have two options:

### Option A – Evaluate all samples at once (merge JSONs)

Our codebase expects a **single JSON** as input for convenience. While we release data in multiple subdirectories for flexibility, you can merge them into one file before running inference.

We provide a merge utility: **/llaso/data/data\_merge.py** (in the anonymized repo). It merges multiple JSON shards into a single JSON.

1.**Extract** the subset:

2.**Edit** the config section at the top of **data\_merge.py** :

3.**Run** the merger:

4.**Run inference/evaluation** on the merged file

### Option B – Quick test on a single JSON

You can also point your evaluation script to **any single JSON** within `audio_text/`, `pure_audio/`, or `text_audio/`.

---

## 6. Reproducibility notes

- The subset preserves **schema, modality configurations, and task definitions** used in the full benchmark and training corpora.
  - Relative audio paths in JSON are resolved against the extracted folder; you can also convert them to absolute paths if preferred.
  - The anonymized repo avoids any links or metadata that could reveal author identity. Full project links and large-scale artifacts will be re-connected after the review period.
-