SANITY CHECK

AUDIOLDM2 300-FULL (FAD)

The experiment was repeated using the whole datasets for two models (AudioLDM2 and Stable Audio Open), in order to ensure that using 300 random samples from each is a reasonable approximation.

STEPS	BASELINE	SUBSET (300 SAMPLES)	FULL SET
10	AUDIOCAPS	0.2198	0.1633
25	AUDIOCAPS	0.1756	0.1151
50	AUDIOCAPS	0.1814	0.1104
100	AUDIOCAPS	0.1809	0.1128
150	AUDIOCAPS	0.1666	0.1122
200	AUDIOCAPS	0.1808	0.1129

STEPS	BASELINE	SUBSET (300 SAMPLES)	FULL SET
10	CLOTHO	0.6052	0.7756
25	CLOTHO	0.5501	0.7267
50	CLOTHO	0.5401	0.7237
100	CLOTHO	0.5472	0.7276
150	CLOTHO	0.5579	0.7429
200	CLOTHO	0.5677	0.735

SANITY CHECK

SAO 300-FULL (FAD)

The tables show the newly computed sanity FAD scores and the original ones (which are computed from the 300-sample subsets, randomly chosen from both baselines)

STEPS	BASELINE	SUBSET (300 SAMPLES)	FULL SET
10	AUDIOCAPS	0.4076	0.3809
25	AUDIOCAPS	0.3003	0.2778
50	AUDIOCAPS	0.298	0.2771
100	AUDIOCAPS	0.2969	0.2706
150	AUDIOCAPS	0.29794	0.2708
200	AUDIOCAPS	0.29791	0.2718

STEPS	BASELINE	SUBSET (300 SAMPLES)	FULL SET
10	CLOTHO	0.5394	0.7308
25	CLOTHO	0.4844	0.6779
50	CLOTHO	0.4861	0.6764
100	CLOTHO	0.4867	0.6723
150	CLOTHO	0.4848	0.6747
200	CLOTHO	0.4853	0.6755

THE RESULTS SHOW A (CONSISTENT) VARIATION IN THE ABSOLUTE SCORE VALUES

CORRELATION

The absolute values change, varying by a somewhat consistent amount given a certain model and baseline. Following this first check, the correlation between the scores of the original experiments and the sanity check were computed to see if the pattern observed for varying inference steps holds.

MODEL	BASELINE	ANCOVA (pvalue)	PEARSON CORR.	P.C. (pvalue)
AUDIOLDM2	AUDIOCAPS	0.9242	0.9455 / 1	0.0044
AUDIOLDM2	CLOTHO	0.9634	0.9615 / 1	0.0022
SAO	AUDIOCAPS	0.9647	0.9985 / 1	0.000003
SAO	CLOTHO	0.9419	0.9937 / 1	0.000058

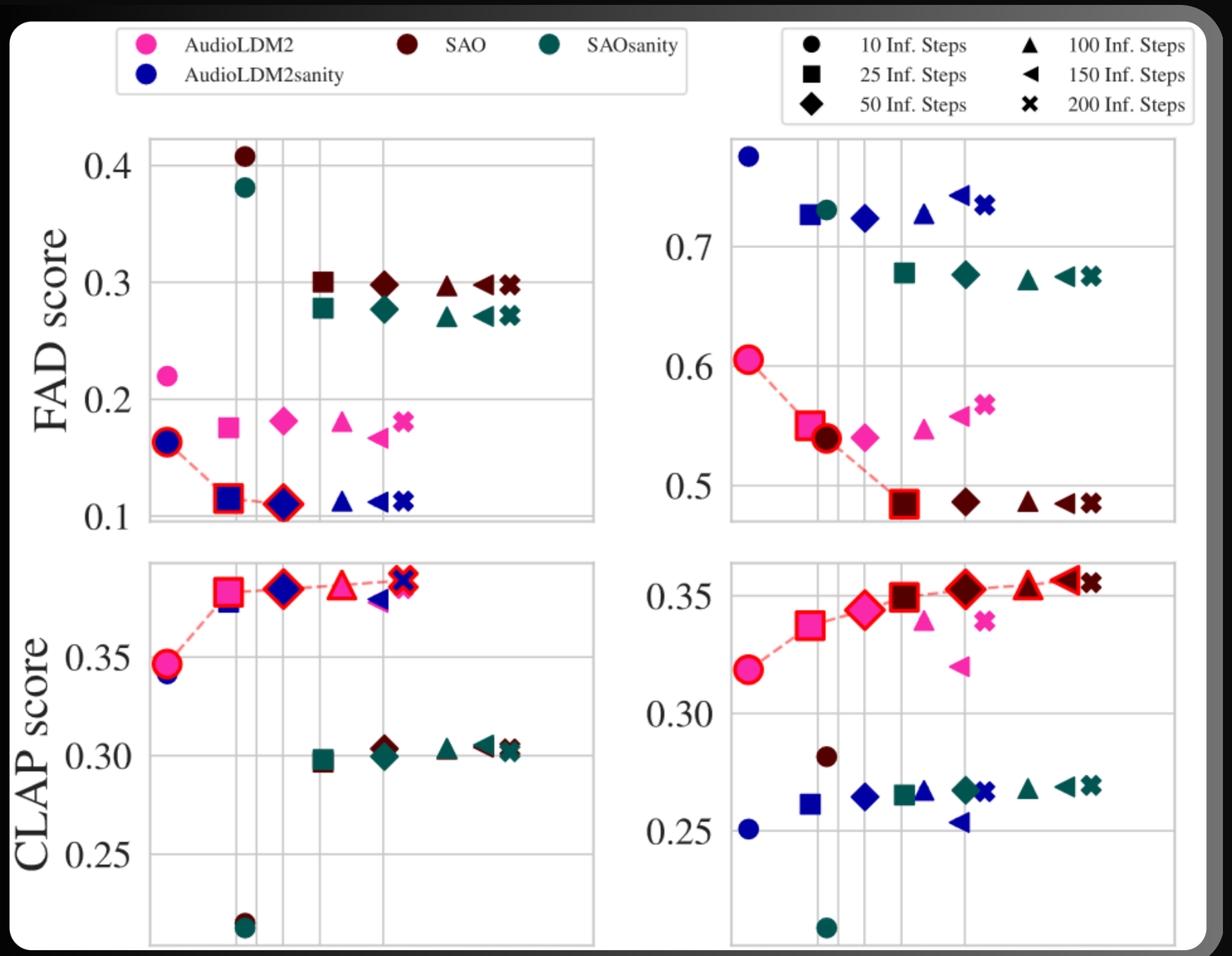
ANCOVA

Determines whether the relationship between steps and score are statistically similar across the two experiments conditions (by checking the interaction term's p-value).

PEARSON COEFFICIENT

Assesses the linear relationship between the two score types (300 and sanity check).
A high coefficient (>0.9) indicates that the two measurements are strongly linearly related. Measured from 0 to 1.

For ANCOVA, the null hypothesis is that the slopes of the two groups are the same. Thus, here a high p-value means that the two sets of values are correlated, meaning they describe the same pattern. For the Pearson Coefficient, a low p-value proves that the observed correlation between the two measures is statistically significant.



To better visualize the similar evolution of the scores as the number of steps changes, this graphs displays the subset and sanity generations like the original plot in slide one.

The pareto plots between the original scores and the sanity check scores are equivalent