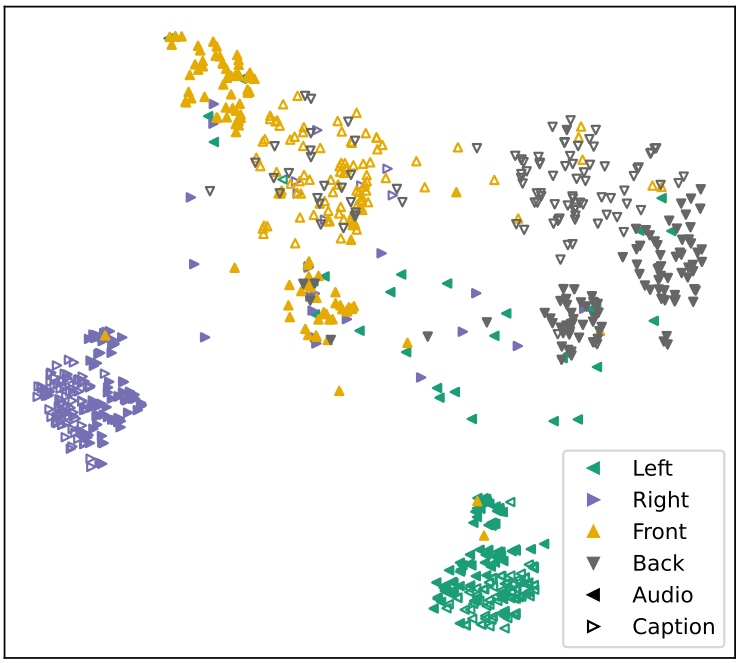
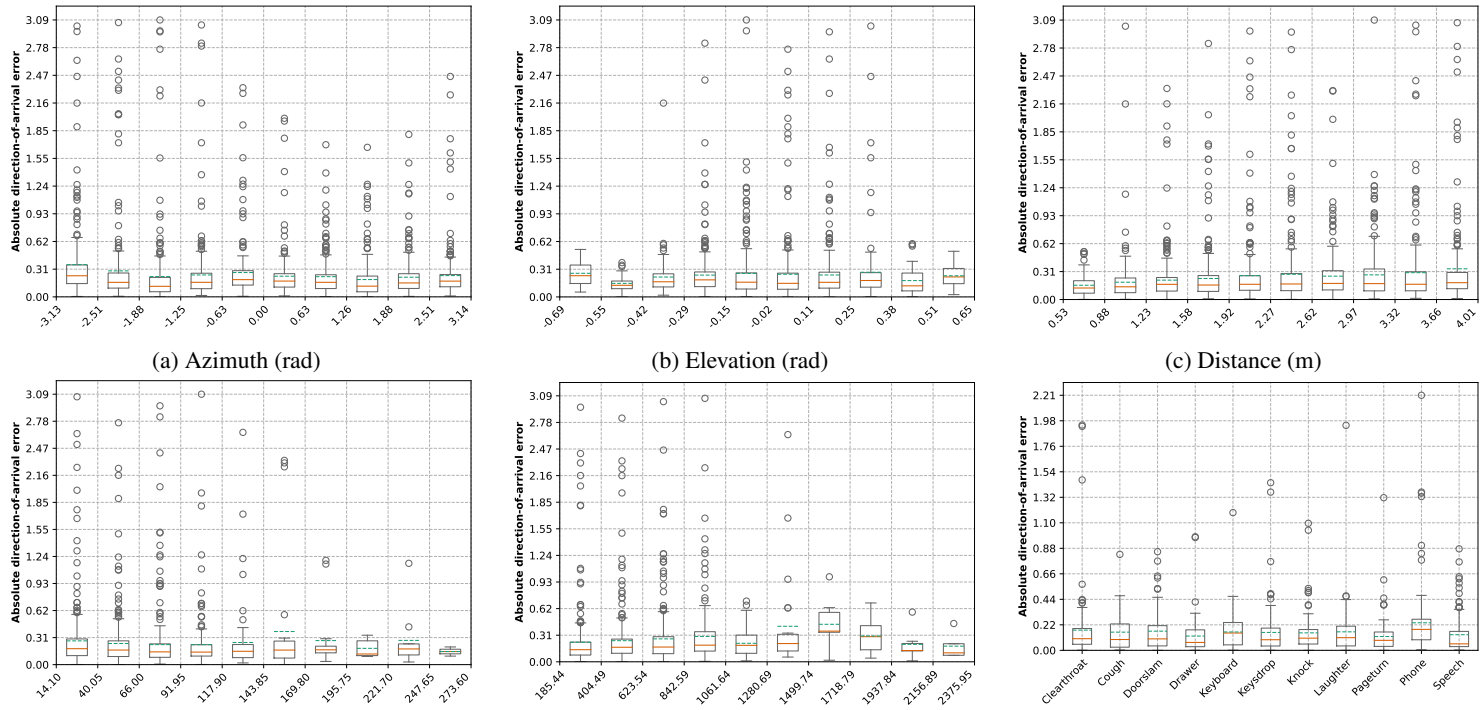


Figure R.1: UMAP projection of ELSA embeddings of the test splits of Spatial-Clotho and Spatial-AudioCaps. Filled markers are obtained from spatial audio, and hollow markers are obtained from spatial captions. The UMAP projection was fitted with the train splits of Spatial-Clotho and Spatial-Audio caps, and we made use of supervised dimension reduction to highlight the direction differences rather than the semantic differences in the embeddings.



(a) Azimuth (rad) (b) Elevation (rad) (c) Distance (m)

(d) Floor area (m^2) (e) Mean T30 (ms) (f) TUT Sound Events 2018 Semantic classes

Figure R.2: Boxplots of absolute direction-of arrival errors predicted by 2-layer MLP. Figs. (a)–(e) show the Spatial Audiocaps and Spatial Clotho test sets errors by different categories. Fig. (f) shows the predictions of the test set of TUT Sounds 2018 by different semantic classes. For all figures, boxes represent the interquartile range, solid orange lines are the median, and dashed green lines are the mean.