---

rebuttal element

Questions on the proposed evaluation metrics.

Comparative diversity score is not penalizing diversity, it is only penalizing differences between the absolute diversity score obtained on generated layouts, and the one obtained on real layouts. Same thing for overlap and alignment. Therefore these comparative scores can indeed be easily compared from different datasets, even with high variation of any given property between and within those datasets (such as the number of layout elements). Our evaluation metrics are not used at all during training, so that evaluation scores remain independent from the training process. The only metric used during training is the binary cross-entropy loss function, which is agnostic to our evaluation metrics. We will add these clarifications in the article and will also provide pseudo-code for our evaluation metrics.

corresponding article update :

- subsection "Applying a same comparative metric on different dataset"
- subsection : "Independence between training metrics and comparative metrics"
- pseudo-code : As planned, a latex version of pseudo code for each metric has been edited but due to lack of space in the article we had to add these pages to the git rather than to the article. These pseudo-codes are yet indicated in section 3.3.

---

rebuttal element

Specific usage and interpretation of these metrics.

We agree with reviewers that adding specific examples and interpretations of our metrics usage could be of great interest, therefore we will use the remaining pages to add our metrics scores, visuals and explanations for the following extreme cases :
- poor training versus intensive training
- training and evaluating on a dataset with low alignment versus training and evaluating on a dataset with high alignment
- same operations for overlap, and diversity

Results on layouts with different number of elements

We already experimented our model on layouts with different number of elements (e.g. 5 or 7 elements) without changing any parameter of the model. Since our paper aims at defining and demonstrating the theoretical value of our new evaluation metrics, we decided to present our model in the simplest possible way, but these complementary results can easily be added to the specific metric usage subsections.

corresponding article update (end of part 3.3)

- section 3.5 : "Additional results : application of the quantitative evaluation metrics on specific examples"
- tables 2, 3, and 4
- figures 4 and 5
- evaluation codes, samples visuals and sub-datasets used in the additional experiments are added available to the git