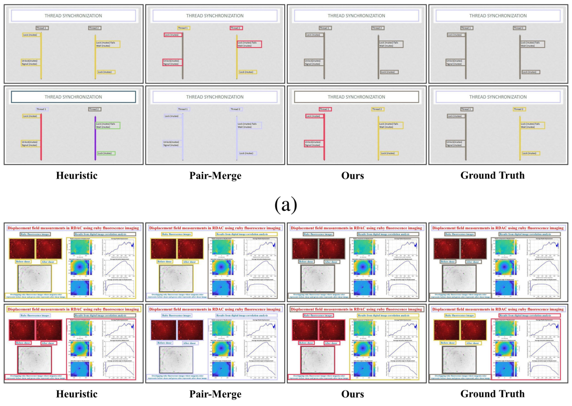
Unsupervised Graphic Layout Grouping with Transformers

1. Case study

We present two examples in Figure 1, which display predictions from different systems along with the corresponding ground truths. For each case, we provide two levels of groupings: coarse-grained in the 1st row and fine-grained in the 2nd row.

In the first case (Figure 1a), we observe that both the Heuristic and Pair-Merge models struggle to perform well when faced with long and thin abnormal shapes. In contrast, our method effectively handles such cases, resulting in improved performance. This indicates that our method is more robust in handling abnormal shapes.

Moving on to the second case (Figure 1b), we can see that both the Heuristic and Pair-Merge models fail to appropriately handle the title and content. They simply combine items that are close to each other, leading to incorrect groupings. In contrast, our method avoids such mistakes. This illustrates that our method has a better capability to handle structural information.



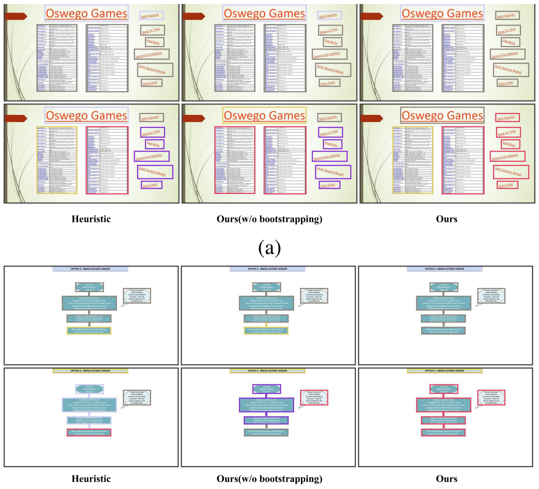
(a)

(b)

Figure 1. Two example slides with grouping predictions from heuristic algorithm, Pair-Merge, our system, and the ground truth. Each slide contains hierarchical groupings with a coarse-level (1st row) and a fine-level (2nd row).

2. Effects of Bootstrapping

To demonstrate the effect of bootstrapping,we alse show two examples with predictions from heuristic algorith, our system without bootstrapping and our system with bootstrapping in Figure 2. We can observe from Figure 2 that methods without the inclusion of bootstrapping (2nd column) exhibit similar errors to the heuristic algorithm (1st column). However, when bootstrapping is incorporated (3rd column), these methods show improvements. This result demonstrates that the addition of bootstrapping during training can enhance model performance and address inherent contamination issues in the training data.



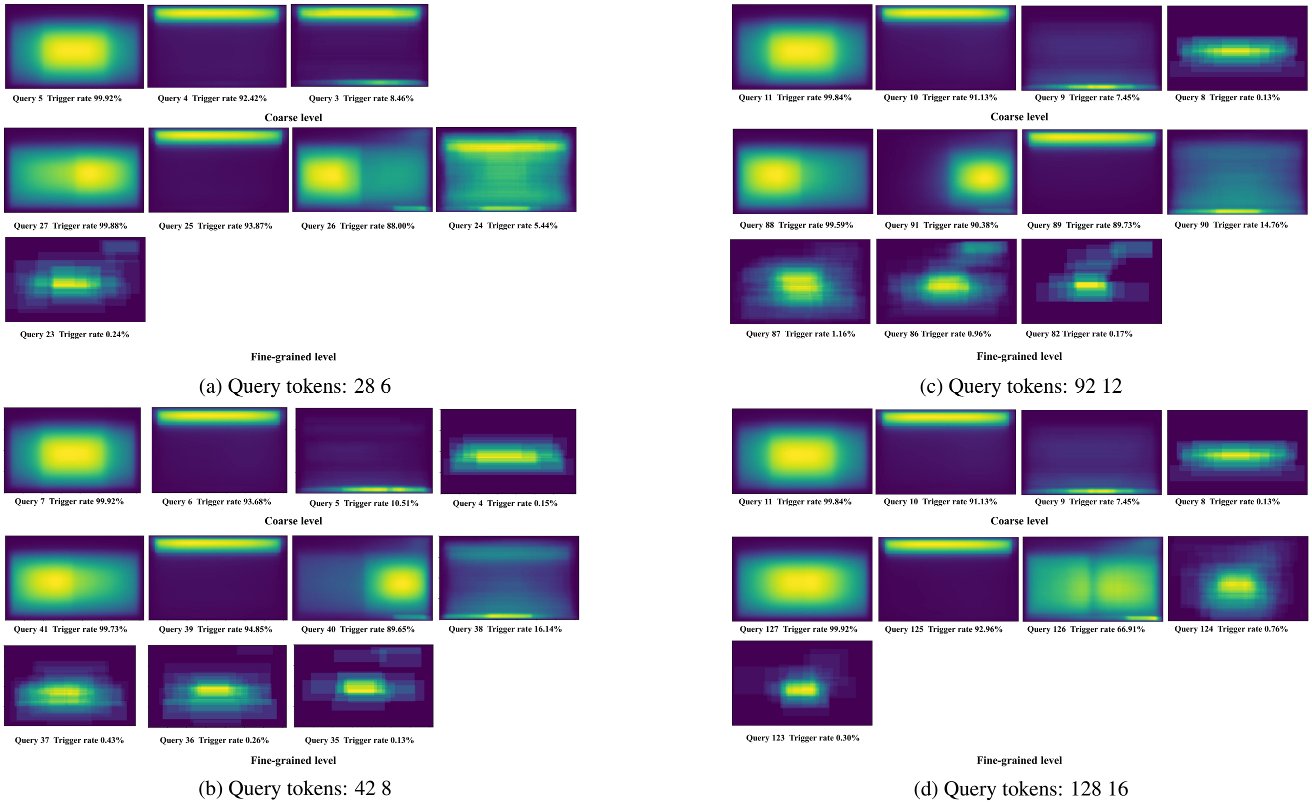
(a)

(b)

Figure 2. Two example slides with grouping predictions from heuristic algorithm our system without bootstrapping, and our system with bootstrapping. Each slide contains hierarchical groupings with a coarse-level (1st row) and a fine-level (2nd row).

3. Triggered group tokens under various group tokens settings

As shown in Figure 3, we evaluate the efficacy of group tokens under various settings by visualizing the predicted bounding boxes generated by these tokens. To this end, we randomly sample 11k slides from the training dataset. However, some group tokens may not have learned meaningful representations and may not correspond to any objects in the dataset. To mitigate this issue, we filter out such tokens that have a triggered rate of less than 0.1% and only visualize the activated group tokens.



(a) Query tokens: 28 6

(c) Query tokens: 92 12

(b) Query tokens: 42 8

(d) Query tokens: 128 16

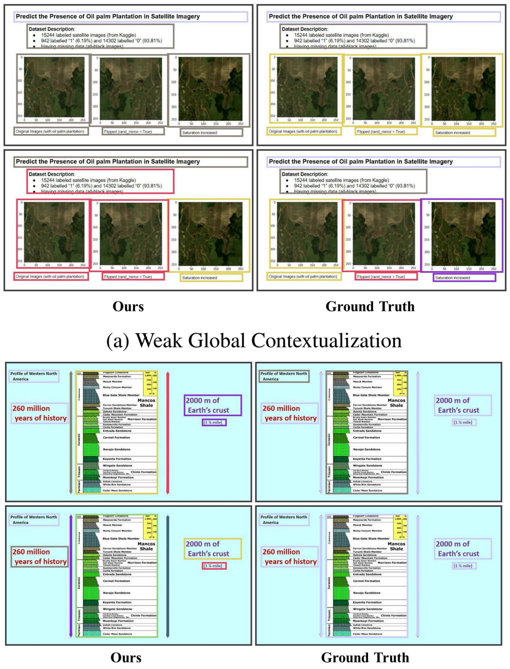
Figure 3. Heatmaps of all box predictions on 11k sampled slides from training dataset under various group tokens settings. We only visualize the triggered group tokens with a triggered rate greater than 0.1%

4. Error cases

We present additional error cases described in section 4.3.4 of the main text.

In the first case (Figure 4a), although our models perform well at the coarse-grained level (1st row), they struggle to accurately group the three horizontal parallelism title + content patterns at the fine-grained level (2nd row).

In the second case (Figure 4b), we can observe that the main picture in the slide contains various notations, such as the statement '260 million years of history' and an arrow shape indicating the timeline. However, our method fails to capture the semantic information and correctly group these elements together with the main picture.


Predict the Presence of Oil palm Plantation in Satellite Imagery
Dataset Description:
Ours
Ground Truth
Profile of Western North America
260 million years of history
2000 m of Earth's crust
Mancos Shale
Ours
Ground Truth

(a) Weak Global Contextualization

(b) Missing semantic content.

Figure 4. Examples of two main error cases