Supplementary: Variational Transformer Networks for Layout Generation

Diego Martin Arroyo[1]
martinarroyo@google.com
[1]Google, Inc

Janis Postels[2]
jpostels@vision.ee.ethz.ch
[2]ETH Zürich

Federico Tombari[1,3]
tombari@google.com
[3]Technische Universität München

1. Attention analysis

The main claim for the effectiveness of our method is its inductive bias towards the relationships between elements. The self-attention layers in our network weigh the relevance of each component regardless of their distance in the input sequence. In this section we analyze the validity of this claim by observing the attention maps on each self-attention layer.

1.1. Encoder

The encoder processes the entire document in a single pass. In the case of PubLayNet, where $n_{\text{heads}} = 4$, we observe that in the first layer elements are independent of each other, since no element receives any attention. In subsequent layers, elements start to consider others in the computation. In fig. 1 we show a visualization of this process.

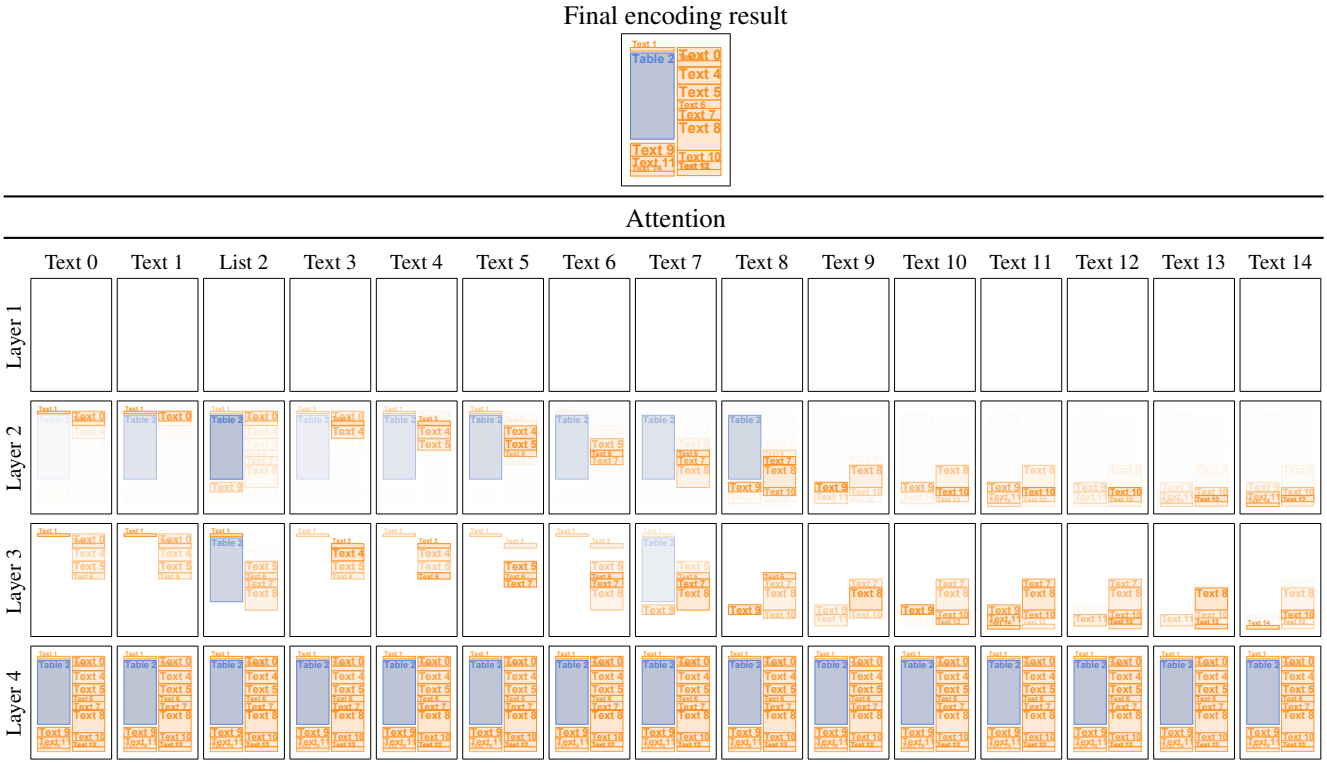

Figure 1: Attention visualization. A higher color intensity reflects a higher attention weight. In the first layer, no attention is paid to any element. In subsequent layers, other elements are considered regardless of their distance in the sequence. This is particularly useful to model two-column documents: for example, in the attention map for Text 8, Table 2 has a significant weight, despite their distance in the input sequence. The network correctly identifies it as a relevant element to consider.

1.2. Decoder

During the autoregressive decoding, the network not only relies on the encoded document vector z to determine the location and size of the next element, but also on the result of the previous iterations via self-attention. In fig. 2 this process is shown.

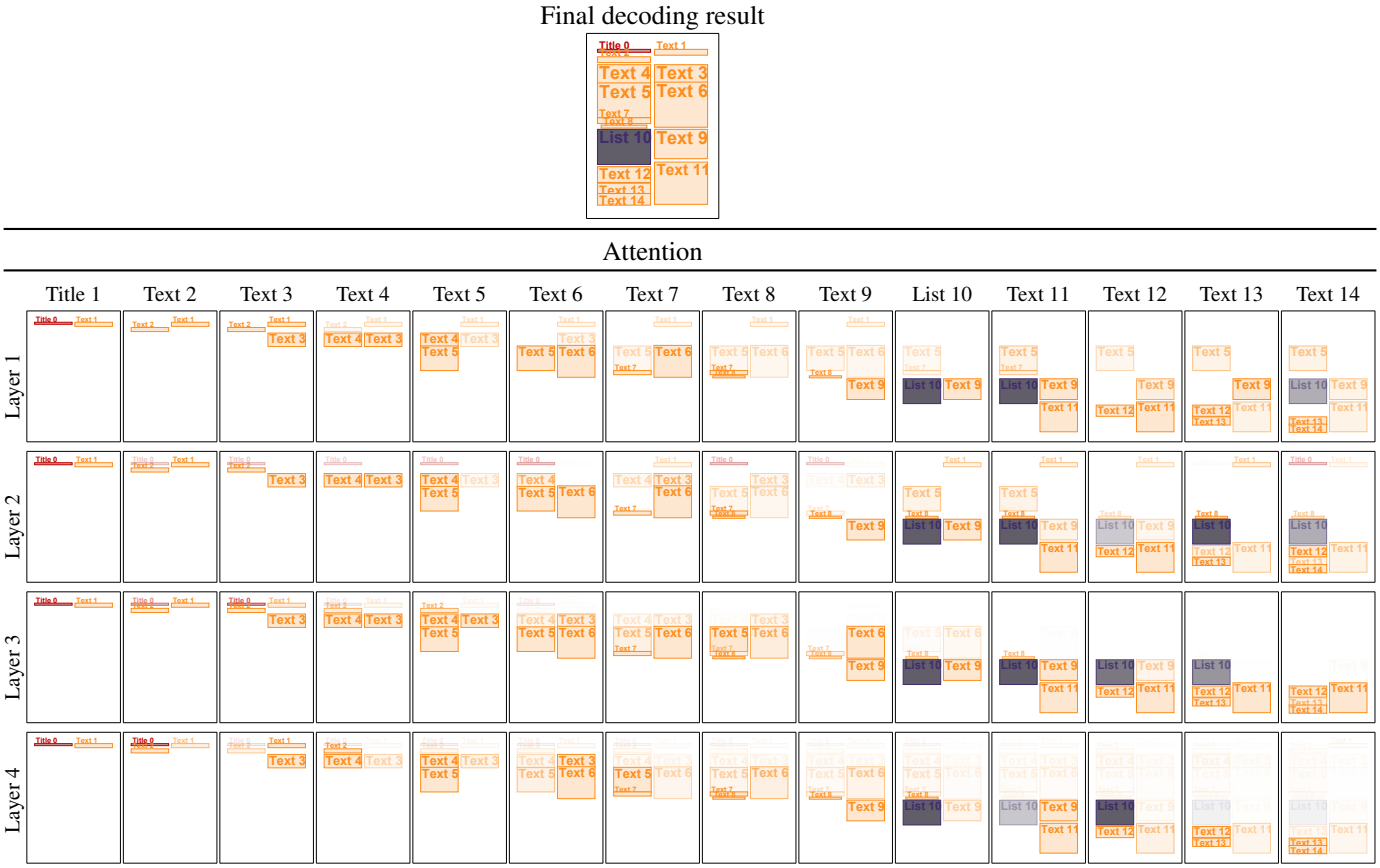

Figure 2: Visualization of the relevance of existing elements during the autoregressive decoding.

2. Latent Space Analysis

The properties of its latent space are an important aspect of any VAE. In this section, we show several experiments to analyze the results of interpolating in latent space as well as the effect of each individual latent vector in the non-autoregressive decoder setting.

2.1. Latent Space Interpolations

In fig. 3 we show linear interpolations between two random vectors $z_1, z_2 \sim \mathcal{N}(0, 1)$ and the intermediate results between them on PubLayNet. While the space is not perfectly smooth (some elements only appear in the intermediate samples), the results are not completely arbitrary, and each intermediate value $z' = z_1 + \lambda \cdot (z_2 - z_1), \lambda \in (0, 1)$ produces a valid output.

Figure 3: Linear interpolations between two vectors in latent space, and evenly-sampled interpolated values.

2.2. Adding/Removing Latent Vectors using Non-Autoregressive Decoder

When training, the non-autoregressive decoder layouts $x \in \mathbb{R}^{l \times d_1}$ are encoded as latent representations $z \in \mathbb{R}^{l \times d_2}$ [1]. This experiment aims to investigate the consistency of these latent representations by removing elements from the latent code element by element - *i.e.* $z \in \mathbb{R}^{l \times d_2} \rightarrow z' \in \mathbb{R}^{l \times d_2 - 1}$. By consistency we mean that removing one latent vector does not drastically change the decoded layout. Qualitative results can be found in fig. 4. We observe that the latent code in fact appears largely consistent. Each new latent code appears to only introduce one new element in the layout while minimally changing the relative arrangement. This also implies that our model could be applied to the task of layout completion.

Figure 4: Investigation of stability of the latent space. From left to right, we add elements to the latent vector. We qualitatively observe that this results in reasonable extensions of the layout.

[1] Technically the encoder parameterizes a distribution over z. However, for the sake of simplicity we shall consider only the mean in this experiment

3. Qualitative results

The small amount of samples presented in the main paper is not capable of conveying the diversity and quality of the synthesized data. In order to provide a clearer understanding, in the following we show a larger amount of samples for each dataset. These results are rendered using the network output coordinates without any postprocessing or cherry-picking applied to them to hide imperfections or failure cases. The first rows show samples from other methods for comparison.

3.1. PubLayNet

Figure 5: Additional synthesized layouts on PubLayNet using an autoregressive decoder and the result of feeding the layout to a document renderer.

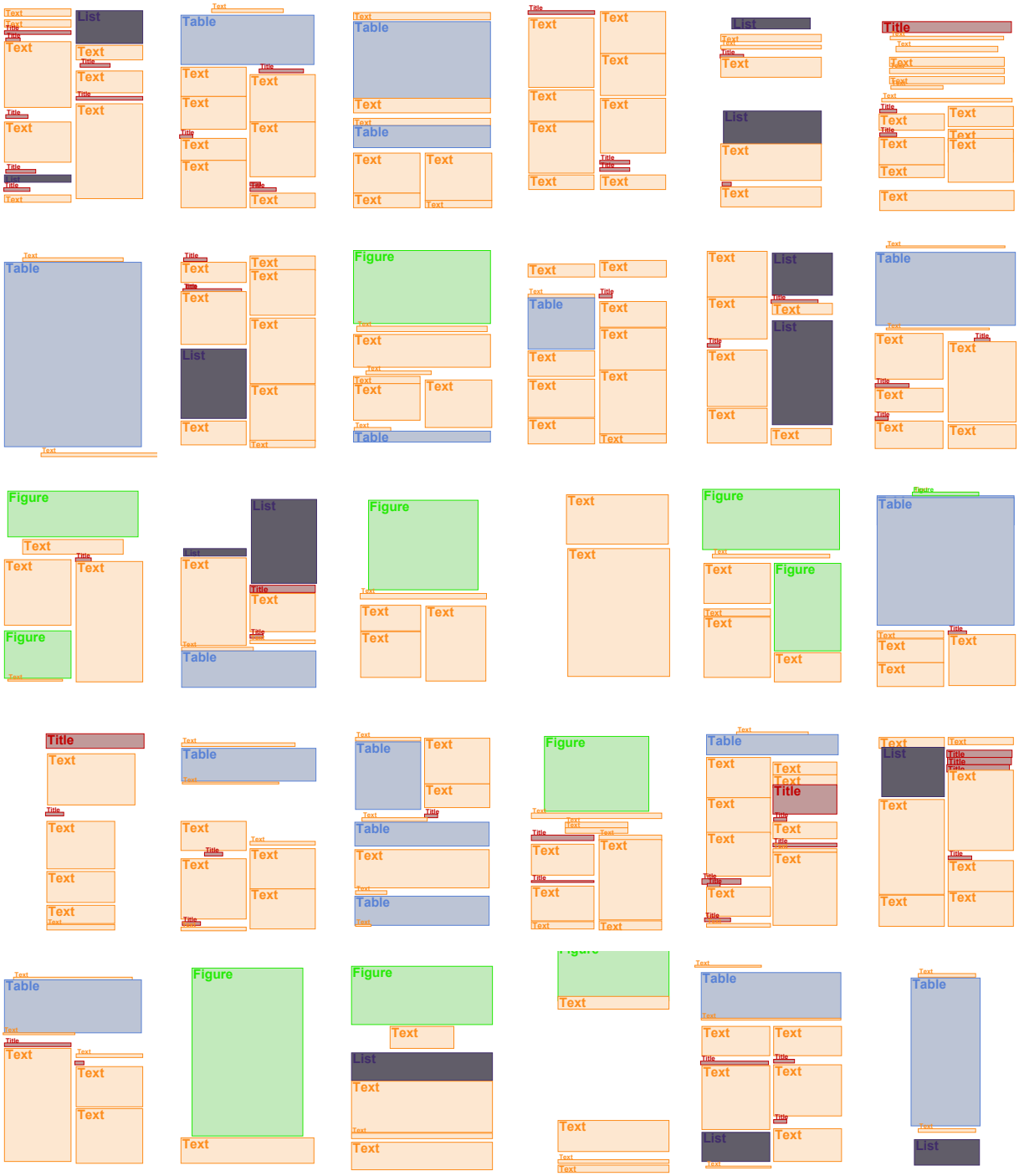

Figure 5: (Cont.) Synthesized layouts on PubLayNet using an autoregressive decoder.

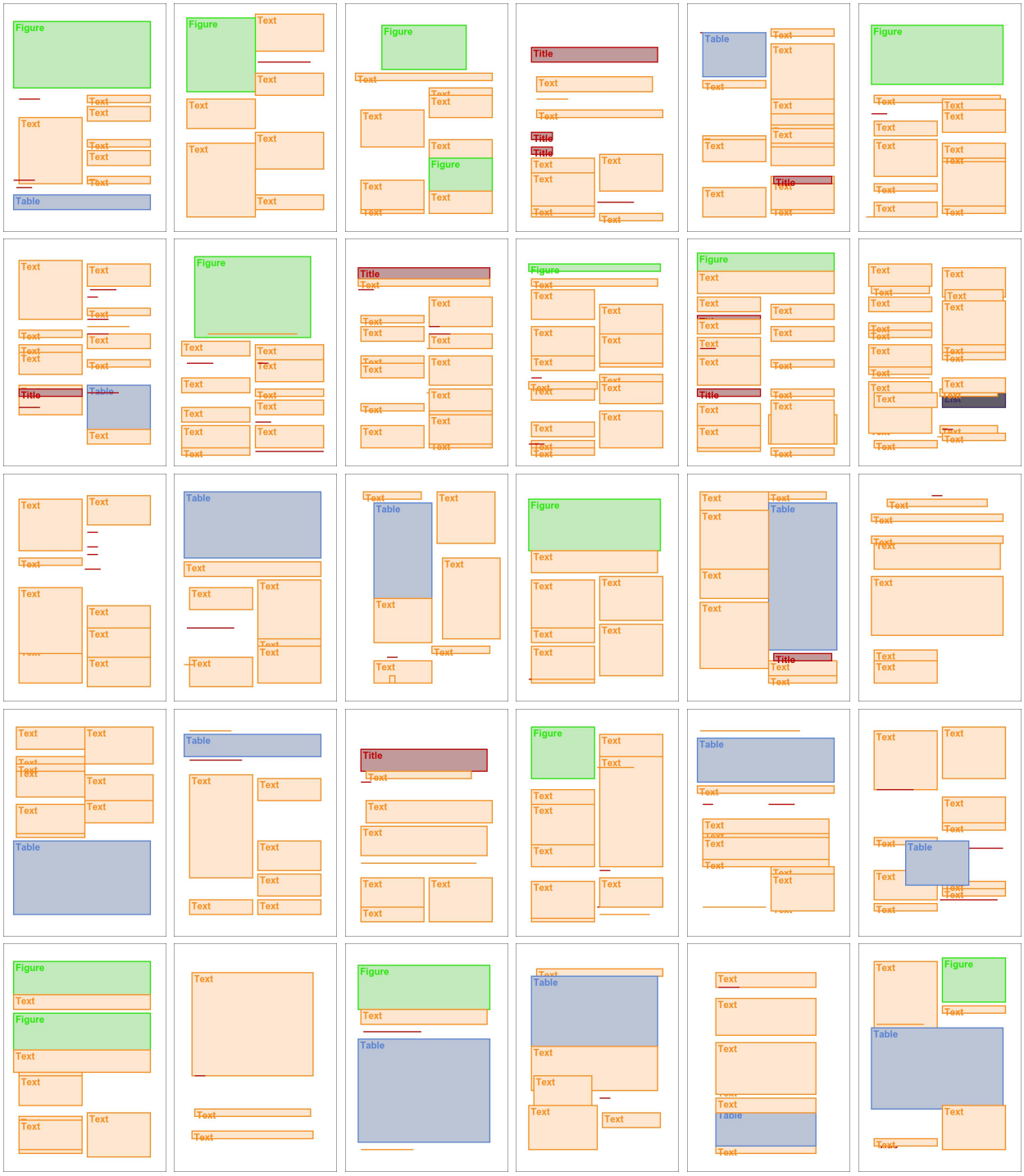

Figure 5: Additional synthesized layouts on PubLayNet using a non-autoregressive decoder.

3.2. RICO

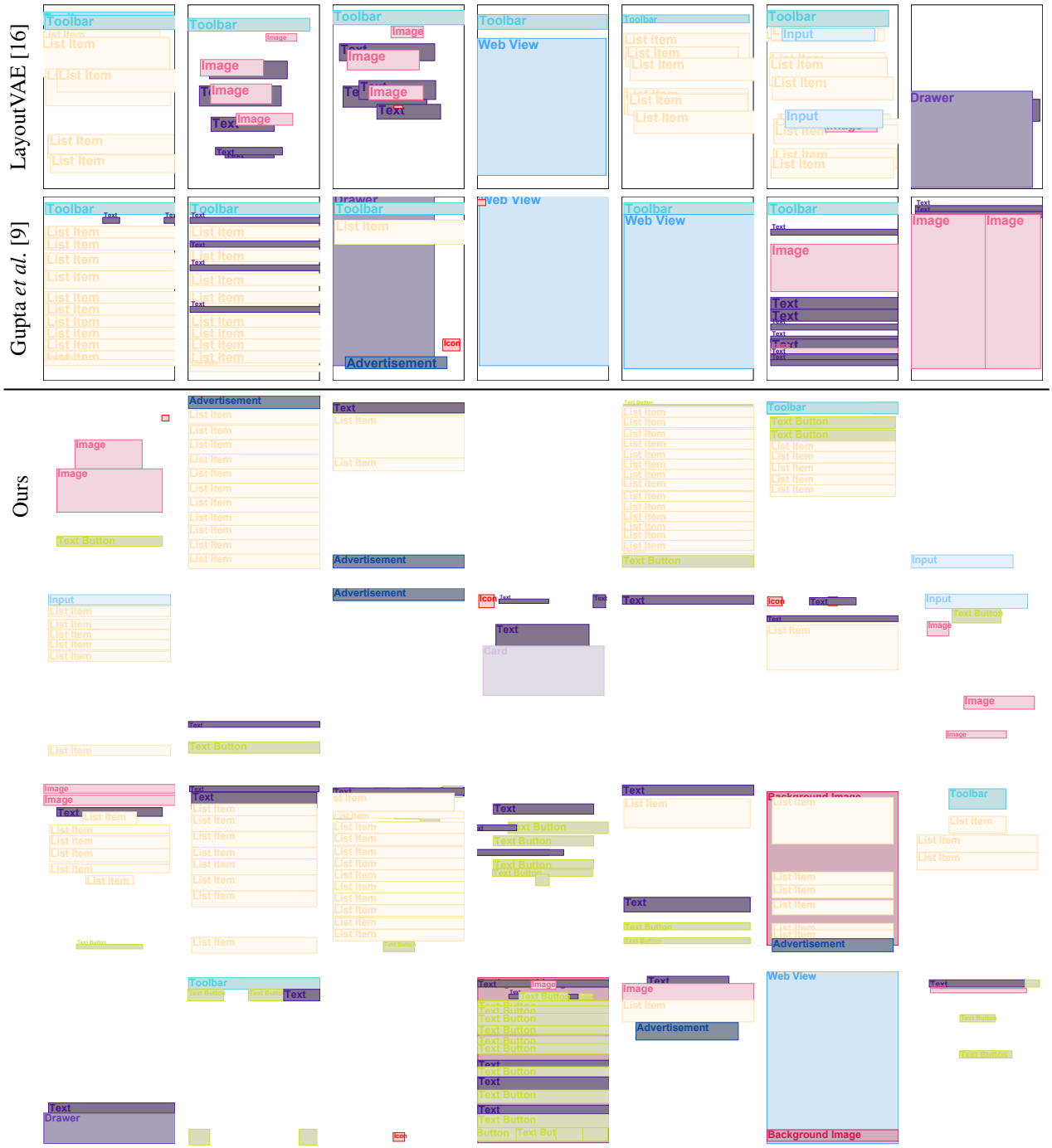

Figure 6: Synthesized RICO examples.

Figure 6: (Cont.) Synthesized RICO examples.

3.3. COCO-Stuff

Figure 7: Synthesized COCO-Stuff examples.

3.4. SUN RGB-D

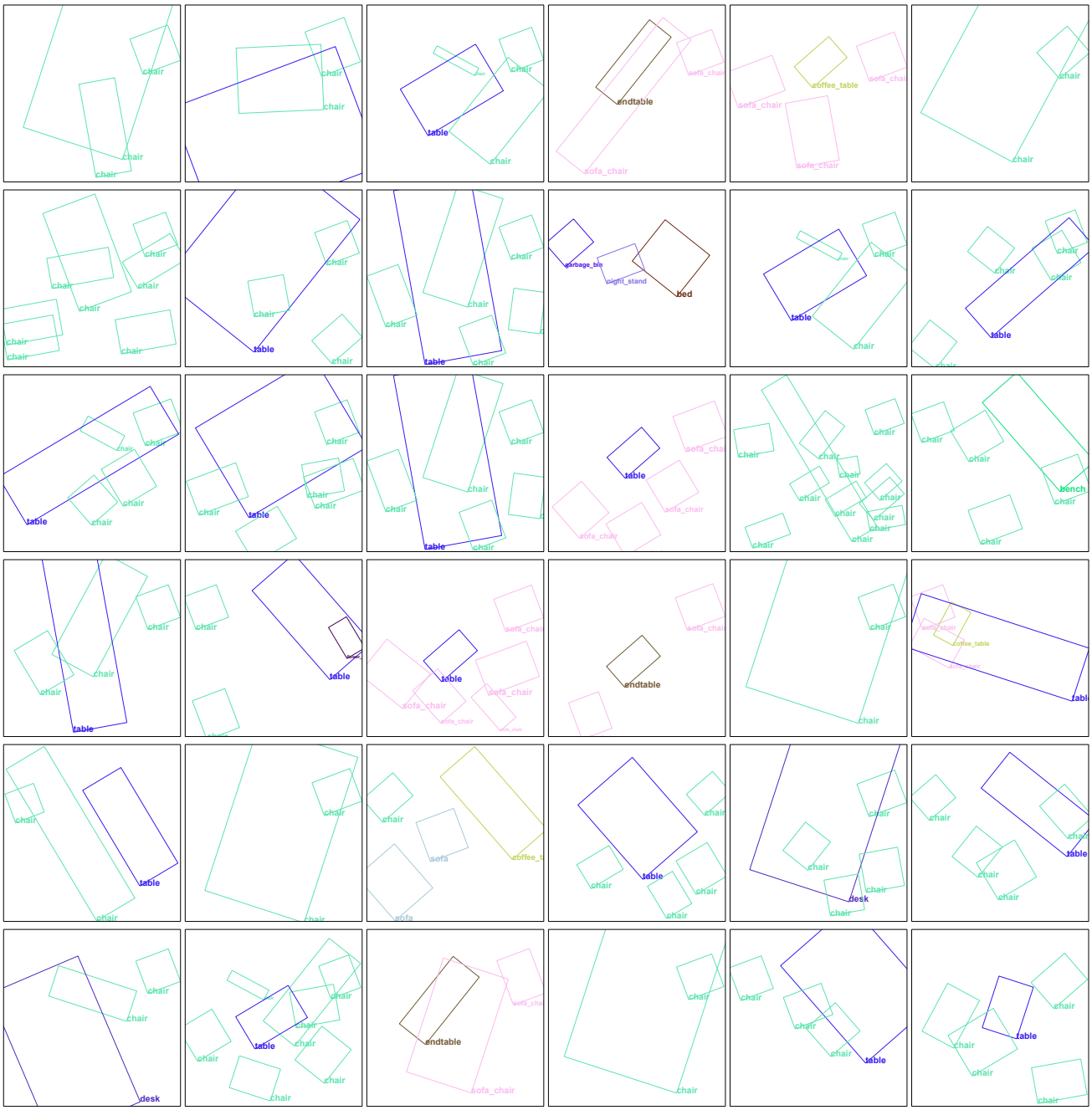

Figure 8: Synthesized SUN RGB-D samples.

3.5. Variational Transformer Networks for Language Modeling

The main motivation for our architectural choices is the success of self-attention in the field of natural language processing. It begs the question, how well does our method perform on text generation tasks? We train our autoregressive model on a dataset of 800K Amazon book reviews from [1] with no major modifications to the architecture: we simply replace the first fully connected layer with a word embedding (learned from scratch) and the output shape to that of our target vocabulary (30K words). For this experiment we use 6 self-attention layers on both encoder and decoder. In tab. 1 we show the results of our method. The reviews have correct grammar and are highly realistic. This shows that our method can be applied to other tasks where the input is a sequence of data.

Real samples	A great read focusing on the main character named Shadow. The gods mentioned in the title are a dark and fading lot, and the story could be seen as a sort of parable about the old making way for the new. Good entertainment, though.
	Just finished this book. I very much liked it. It kept me engaged throughout the entire reading. It sometimes seems as if you're not fully immersed in Shadow's like and persona, however, at the end of the book, I felt as though I knew enough to feel satisfied.
	Long winded and no excitement. It just plodded. On until it ended. complete waste of time, do t plan on reading anymore of this author's material
	This is a really imaginative story that takes you totally by surprise and all over several dimentions.
	Although I spent most of this book being not quite sure what was going on, once I finished it and took the whole thing in I loved it. The revelations at the end are what really sealed the deal for me, finding them very clever and fulfilling.
	Great book. It felt as though the book moved both slowly and quickly as only a true classic can.
	Great book! I love the Blossom Street Series! I am hooked on that series! Got the book in a reasonable amount of time! Very happy with the book! Came in great condition!
	I have enjoyed Debbie Macomber's other book series so I decided to try this Blossom a Shop series. I love it and lose myself in the lives of the characters!
	Loved it.
	This series is awesome as good as all her books .I really enjoyed these books a lot
	Anyone who enjoys reading about women's friendships will love this book. As a knitter I really like reading about the knitting and have taken away some ideas to use in my knitting projects, especially for the charitable knitting group I'm in!
Ours	interesting ride of danger and suspense i think clancy has been on our lookout to clean ways
	these books are awesome and can't wait to read more about the history of the city going through some presidential topped
	a really good novel i couldn't put down there are many ups and downs yet so uplifting
	very good book about an amazing woman this one leaves you begging and be entertained
	therapy and exciting parts should be read by everyone real life persons an approach thanks
	thanks for opening my eyes to everyone i've never read anything like it was very great
	a great read one learns and never boring facts in school so that you found it very believable
	a few of the characters were not particularly compelling but they never got any better it let me down with many people
	good suspense and great writing will open a person's eyes on other subject matter
	great read and interesting ending chapters are sad and far away from you
	very interesting read hard to put down lehane does have really researched his subject matter but he made it realistic plus the ending in a quit
	we learn about africa and from this novel every time we really need a man and makes it
	a must read for any real good reason it's been years since the author came to life of your face
	great book just like divergent and always worth reading this year messages of speech has been thorough
	love it all my life truly a page turner and what else is happening great stuff about endurance
	excellent end to an amazing trilogy energy and suspense can't wait there is another life
	good read but takes forever to write about things and its our real lives
	another success for john o'donohue and the same day i meet him but excellent a lot of true emotions
	unexpected turn in surprises and surprise ending was a UNK think again and again about how worthwhile
	a great read and i love his writing and it made me think about leader of tennis authorities
	yes it is thinking me things i never knew about glad you were written or heard from author
	good read makes you think and feel how it was affecting the people throughout every new books

Table 1: **Top**: Real Amazon book reviews from [1]. **Bottom**: reviews generated with our method.

4. Convergence tests

In order to determine the required number of elements for our model to generalize, we train our autoregressive model on various subsets of the PubLayNet training data, and evaluate the number of unique DocSim matches on 1000 samples. We average the results across 5 identical trainings. In fig. 9 we show the results. For this particular dataset, 50K training samples are enough to generalize well.

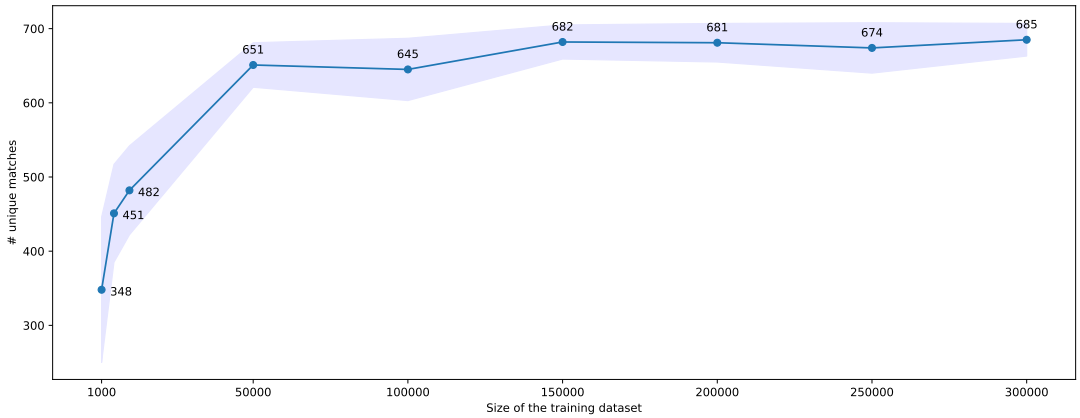

Figure 9: Convergence results on PubLayNet $\pm 1\sigma$.

5. Distribution analysis

As an additional metric for the ability of our method to capture the layout distribution, in fig. 10 we show the frequency of each location as the center of a bounding box on 1000 samples of the PubLayNet test dataset and our model.

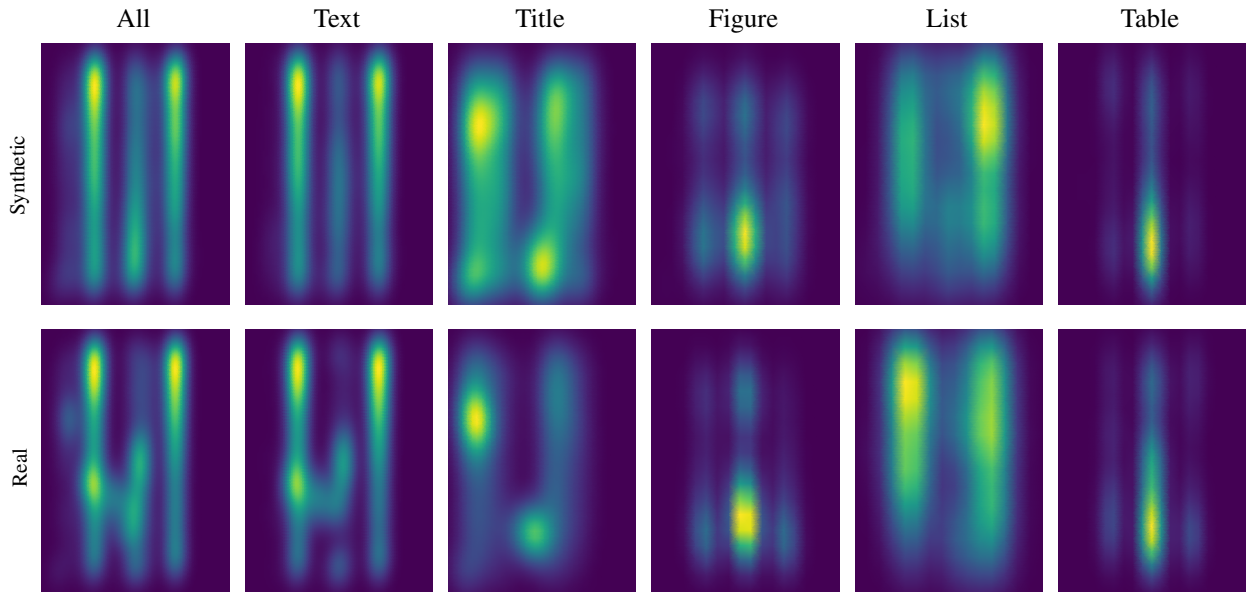

Figure 10: **Top**: distribution of the bounding box center for synthetic data. **Bottom**: real data.

References

[1] Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.