Sketch-to-Layout: Sketch-Guided Multimodal Layout Generation

Supplementary Material

1. Implementation Details

1.1. Training details

For all the analysis and experiments described in the paper, the model has been trained for 10 epochs with a batch size of 128, freezing the ViT and using a cosine learning rate scheduler [2]. The learning rate has been set to 10^{-4}, and no dropout is used.

During training, the order in which the assets appear in the input textual prompt and the order in which they are fed to the vision encoder are randomized, therefore not matching how they are listed in the output. This serves the specific purpose that the model should learn how to relate each element to the others based on their (image or textual) content, without exploiting any deterministic rule mapping the elements listed in the input to their position in the output.

1.2. Data Pre-processing

We performed several pre-processing steps on the three public datasets used in our experiments. First, we crop the content of each bounding box and use an OCR model to extract text content from it. For SlideVQA, we use a large-hole inpainting model to extract the background as a separate asset after masking all foreground bounding boxes. This allowed us to obtain the content necessary for our content-aware experiments. Then, using the same OCR model we extract the font size and font color of text elements and perform data smoothing of these outputs as a post-processing step. This allowed us to have more accurate rendering for debugging and demonstration purposes. The extracted font size was also used in the synthetic sketch generation pipeline as a clustering attribute.

1.3. Synthetic Sketch Generation

To store collected primitives, we use KD-Trees [3] but we achieve similar results qualitatively by sampling from the top 10 closest elements iterating over the full dataset of primitives or by sampling at random from pre-computed centroids using K-Means on the training data. KD-Trees have the advantage to not require pre-computation of centroids and are faster than iterating over the full dataset of primitives (log vs linear complexity).

2. Comparative analysis of the sketch as a guidance method

The same experiment done on PubLayNet was performed on DocLayNet and SlideVQA. We report the results below.

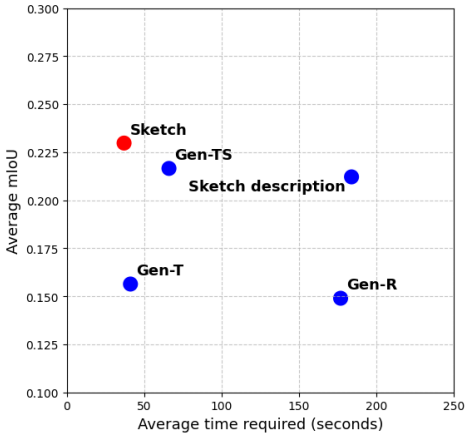

Figure 1. Time-performance trade-off between guidance methods on the DocLayNet dataset.

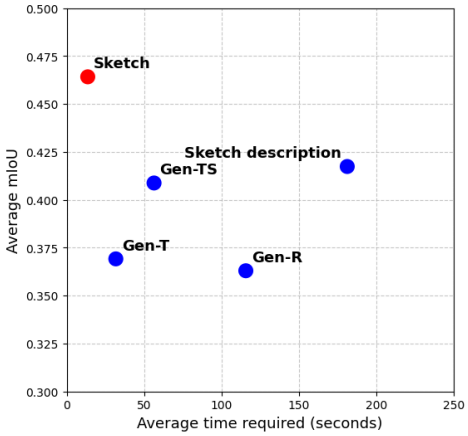

Figure 2. Time-performance trade-off between guidance methods on the SlideVQA dataset.

3. Prompt Examples

For Gen-T, Gen-TS, Gen-R we follow the LayoutPrompter [1] method and use the same prompts, the only difference is that our output format has a JSON structure.

3.1. Textual Description of the Sketch To Layout

The task is to generate a layout based on the textual description of the sketch. We determine whether the sketch is an efficient modality by comparing the results of sketch-based generation to the ones conditioned on a textual representation of the same sketch.

3.1.1. Textual description Creation

Firstly, to create meaningful representations of the sketch, we few-shot prompt Gemini to generate a textual description of the sketch. The prompt is the following:

```
You are an AI assistant, you are given a
sketch made of ink of a layout that could
be a scientific document, a slide or an
ad, the sketch represents the position of
the assets in the real layout. Images are
represented by rectangles with the two
diagonals drawn, whereas text (so titles,
paragraphs, etc.) is represented by ink
lines.

Please give a detailed and quantitative
description of the sketch so that a human
could reproduce the layout only based on
your textual description. You must be as
exhaustive as possible describing all the
elements in the sketch. Never mention that
an image is represented by a rectangle
with two diagonals, directly say its an
image.

Start with describing the overall
structure of the layout such as if it is
two column format, Then, describe the
positions of the image assets and then
text assets one by one using terms like
middle, corner, upper, lower, left, right.
For each asset, describe the size of the
asset with ratios compared to the whole
layout. Your description should be very
detailed. You will be provided with the
exact number of assets, and in particular
with the exact number of text assets and
image assets.

Input:
The sketch has 8 text assets and 2 image
assets.
<<SKETCH>>

Output:
The sketch has 10 assets in total: 8 text
assets and 2 image assets.

The sketch is organized on two columns
that cover the whole height of the sketch,
each column width is half of the whole
width of the sketch.

The following is a description from top to
bottom, left to right of the whole sketch:
there is a small text asset on top of the
first column, right below it there is an
image asset that occupies 1/4 of the
height of the column and the whole width
of the first column. Below it there are 3
text assets that occupy the remaining 3/4
of the first column. The second column has
on top a small text asset, below there is
an image asset that occupies 1/4 of the
height of the second column, below there
are 3 text assets that cover 3/4 of the
second column.

<<OTHER FEW SHOT EXAMPLES>>

Input:
The sketch has N text assets and M image
assets.
```

3.1.2. Gemini Few-shot prompt

After we have obtained few-shot descriptions of sketches for our support samples, we can create the few-shot prompt to query Gemini on Text-to-layout task:

```
Please generate a layout based on the
given information. You need to ensure that
the generated layout looks realistic, with
elements well aligned and avoiding
unnecessary overlap.

Task Description: generation conditioned
on given textual description of the layout

Layout Domain: slide layout

The sketch has 7 assets in total: 5 text
assets and 2 image assets.

The sketch represents a slide with an
image asset acting as background covering
the whole width and height of the slide.

This is a description from top to bottom
of the whole sketch. At the top left part
of the sketch, there is a text asset,
covering 1/4 of the sketch width and 1/4
of the sketch height. Next to it, on its
right, there is another text asset,
covering 1/4 of the sketch width and 1/3
of the sketch height.

Then, at the bottom left, there is a text
asset, covering 1/2 of the sketch width
and 1/2 of the sketch height. Next to it,
on its right, there is another text asset,
covering 1/4 of the sketch width and 1/3
of the sketch height. At the bottom right,
there is a text asset, covering 1/4 of the
sketch width and 1/8 of the sketch height.
```

```
At the bottom left corner, there is an
image asset, covering 1/8 of the sketch
width and 1/8 of the sketch height.
Element Type Constraint: background |
image_0 | page_text_0 | page_text_4 |
page_text_3 | page_text_1 | other_text_2

{
  "elements": [
    {
      "name": "background",
      "bbox": {
        "width": 1000,
        "height": 1000
      }
    },
    {
      "name": "image_0",
      "bbox": {
        "xmin": 18,
        "ymin": 891,
        "width": 86,
        "height": 91
      }
    },
    {
      "name": "page_text_0",
      "bbox": {
        "xmin": 282,
        "ymin": 92,
        "width": 233,
        "height": 237
      }
    },
    {
      "name": "page_text_4",
      "bbox": {
        "xmin": 471,
        "ymin": 504,
        "width": 286,
        "height": 245
      }
    },
    {
      "name": "page_text_3",
      "bbox": {
        "xmin": 51,
        "ymin": 512,
        "width": 387,
        "height": 258
      }
    },
    {
      "name": "page_text_1",
      "bbox": {
        "xmin": 535,
        "ymin": 94,
        "width": 393,
        "height": 278
      }
    },
    {
      "name": "other_text_2",
      "bbox": {
        "xmin": 732,
        "ymin": 893,
        "width": 242,
        "height": 71
      }
    }
  ]
}
```

3.2. Sketch To Layout Gemini

To correctly perform few-shot prompting using Gemini, we define two different input formats depending on whether the content has to be included and given as input to the model.

3.2.1. Sketch-Only to Layout

To generate the prompt given to the model, we leverage 32 support examples randomly selected each time the model is queried. After providing an initial instruction describing the purpose of the task, we provide a specific set of information for each support sample: the type of layout (slide or document), the description of the primitives used to draw the sketch, the names of the assets appearing in the result, the corresponding sketch and its protobuf representation. The following is an example showing how a DocLayNet sample is leveraged when using it as support:

```
Please generate a layout based on the
given information. You need to ensure that
the generated layout looks realistic, with
elements well aligned and avoiding
unnecessary overlap.

Task Description: generation conditioned
on given element types and sketch

Layout Domain: document layout.

To generate the layout you must follow the
sketch represented in the next image,
where each image asset is represented by a
crossed rectangle, whereas text assets
(titles, paragraphs, descriptions, ...)
are represented by straight or wavy
horizontal lines, in particular each
cluster of straight horizontal lines (that
could contain any number of lines starting
from 1) represent one text asset.
Element Type Constraint: picture 0 |
picture 1 | picture 2 | text 3 | text 4 |
text 5
```

The instruction is then followed by the sketch, in image format, and the protobuf representation. As we are working in the sketch-only setting, no information about the assets' content is provided, and only their names are listed. The way the assets are listed and the information are encoded is equivalent to what has been done for the textual baseline, in order to fairly compare the validity of the sketch.

3.2.2. Sketch with Content to Layout

Differently from what has been described before, it is now necessary to include the content of each asset in the prompt. Additionally, such a baseline is used to better measure the performance of Content-Aware PaliGemma. Therefore, for a fair comparison, we use the same input format. For each sample used for support, the prompt is as follows:

The text, which contains the content of textual elements given the content-aware nature of the approach, is then followed by the sketch and the output in protobuf format. While image assets for the support samples are not provided in order not to increase the length of the context too much, those belonging to the sample to evaluate are added and appended immediately after the sketch.

3.3. Layout Prompter Details

```
Please generate a layout based on the
given information. You need to ensure that
the generated layout looks realistic, with
elements well aligned and avoiding
unnecessary overlap.

Task Description: generation conditioned
on given element types

Layout Domain: slide layout

Canvas Size: canvas width is 160px, canvas
height is 120px

Element Type Constraint: background 0 |
figure 1 | page_text 2 | title 3

Asset Contents:
background 0:
<PIL.PngImagePlugin.PngImageFile image
mode=RGB size=1024x768 at 0x7111DF0E1310>
figure 1:
<PIL.PngImagePlugin.PngImageFile image
mode=RGB size=1010x607 at 0x7111DF0BBD90>
page\_text 2: Journey Map
title 3: UX LX CONFERENCE JOURNEY
<html>
<body>
<div class="canvas" style="left: 0px; top:
0px; width: 160px; height: 120px"></div>
<div class="background" style="index: 0;
left: 0px; top: 0px; width: 160px; height:
120px"></div>
<div class="figure" style="index: 1; left:
2px; top: 13px; width: 157px; height:
94px"></div>
<div class="page\_text" style="index: 2;
left: 8px; top: 9px; width: 13px; height:
2px"></div>
<div class="title" style="index: 3; left:
26px; top: 7px; width: 66px; height:
3px"></div>
</body>
</html>
```

3.4. Sketch to Layout Content-Aware PaliGemma

As explained in the main section, the model is given both textual and image assets information in the input, in order to guide the generation. The following is an example of prompt used when training and evaluating out content-aware solution.

```
Please prepare a width: 1700 x height:
2200 layout for the following assets:

text7: Fig. 2 shows the time course
changes in normalized rmsEMG of m.MG,
m.LG, and m.SOL. The rmsEMG in those
muscles increased similarly with
increasing exercise intensity. The rmsEMG
of m.MG for each of the first 30 s at 20%,
30%, 50%, 60%, 70%, and 80% MVC differed
significantly from that during the 30 s of
exercise immediately before (i.e., prior
intensity) (p < 0.05). Throughout the
exercise, the change in rmsEMG of m.MG was
largest in the three muscle groups.;

text5: Fig. 3A shows the time course of
changes in intramuscular pH. We found that
pH was relatively constant, from resting
values (7.06  0.01) until 60% MVC (7.04
0.08), but it decreased significantly (p <
0.05) at 70% MVC and with exercise
progression, being 6.78  0.22 at the end
of exercise.;

text3: Fig. 3B shows the time course
changes in intramuscular PCr. We found
that there were significant differences
after the last 30 s at 40% MVC when
compared with the value obtained during
the first 30 s at 10% MVC (p < 0.05), and
that PCr decreased with progression of
exercise. Above 70% MVC, the values were
significantly different when compared with
those obtained during the 30 s of exercise
immediately before. A linear regression
line was drawn to obtain the highest
```

```
correlation coefficient above the last 30
s of 40% MVC, at which significant
difference was;

text0: Division of data analysis (30s).;

text1: course changes in each parameter,
and Fisher's PLSD post hoc comparisons
were used to determine the significance of
differences of each parameter every 30 s.
A linear regression analysis was used to
examine the relationship between each
parameter. P < 0.05 was defined as
statistically significant.;

text2: 02mus measurement (6 s; once per
three contraction phases).;

text6: Figure I Procedure for data
analysis. Each parameter was analyzed
every 30 s. Muscle phosphocreatine (PCr),
inorganic phosphate (Pi), pH, estimated
ADP and free energy of ATP hydrolysis
(AGATP), pulmonary oxygen uptake (V02pul),
and electromyogram (EMG) were averaged
over 30 s. The data for muscle oxygen
consumption (VO2mus) were obtained during
the third (20-26 s) and sixth (50-56 s)
contractions at each intensity. The V
02mus value of the third contraction was
used to represent the first 30 s of each
minute, whereas the V 02mus value of the
sixth contraction was used to represent
the last 30 s of each minute.;

title4: Results;

figure0 (width: 1386 x height: 765):
<image>. The output should be a single
sentence, in protocol buffer debug string
format.
```

When running our ablation study assessing the usefulness of adding the assets' content to the input, we avoid including text contents and images to the prompt, as the only considered visual input is the sketch. Therefore, only text elements are included, reporting their dimensions but not their content.

4. Content-Agnostic vs Content-Aware Results

Incorporating the content of the assets in addition to the sketch helps the model to better place the assets, especially in cases where the positions of the assets are correct but the order of them is incorrect. Such an example can be seen in Figure 3 where the content-agnostic placement was incorrect due to the misorder of the elements.

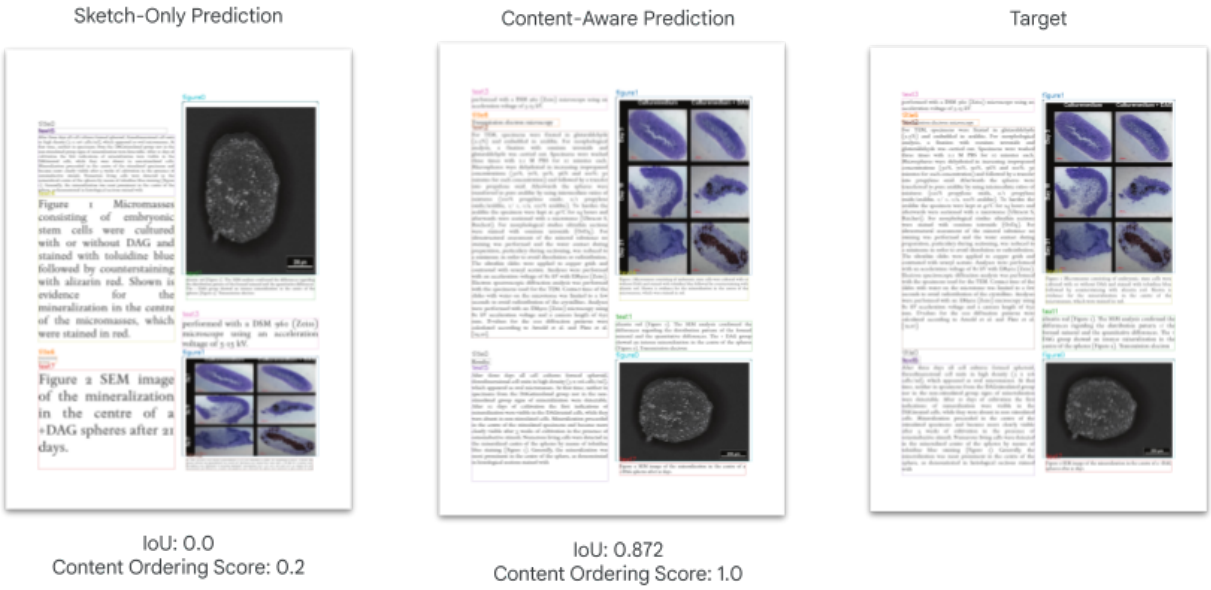

Figure 3. Providing additional assets information helps the model better generate the desired layout.[2]

5. Complete Partial Sketches Results

The results for partial sketches ablation study on all the datasets can be seen in Figure 4. It can be observed that increasing the coverage yields better results, confirming the value of sketch as a guidance prior. However, this increase is not monotonic as can be seen in the increase from 75% to 100% on PubLayNet and 0% to 25% on DocLayNet.

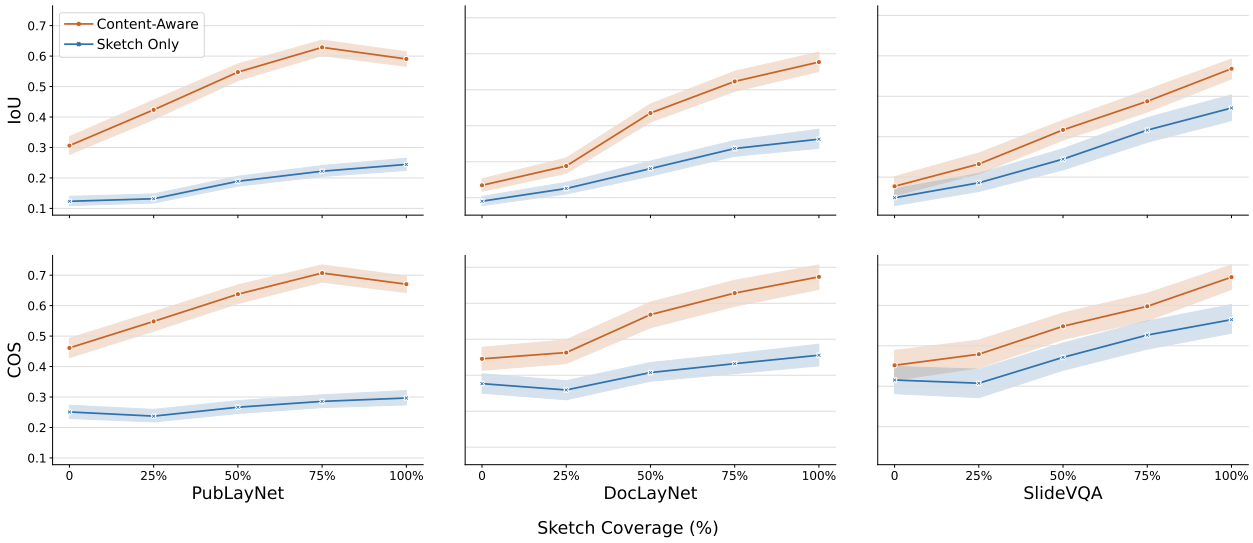

Figure 4. Partial sketch results on all datasets.

6. Synthetic vs Real sketches

The complete results on synthetic and real sketches can be seen in the Table 1 below. The alignment and the overlap metrics of the original layouts are also given in the last two columns, which can be interpreted as reference values that good layouts would have similar values to. There is no statistically significant difference between the metrics for the synthetic and human collected sketches, which confirms that the synthetic sketches are similar to actual sketches.

Dataset	Setting	mIoU	IoU	Overlap	Alignment	COS	Alignment Target	Overlap Target
DocLayNet	Human sketches	0.590 ± 0.171	0.457 ± 0.252	0.003 ± 0.007	0.003 ± 0.0074	0.665 ± 0.296	0.003 ± 0.008	0.0001 ± 0.001
	Synthetic sketches	0.592 ± 0.164	0.466 ± 0.245	0.009 ± 0.031	0.003 ± 0.007	0.669 ± 0.298	0.003 ± 0.007	0.0001 ± 0.001
PubLayNet	Human sketches	0.761 ± 0.132	0.623 ± 0.232	0.003 ± 0.006	0.0003 ± 0.0009	0.699 ± 0.253	0.0002 ± 0.0005	0.0004 ± 0.001
	Synthetic sketches	0.806 ± 0.117	0.675 ± 0.216	0.005 ± 0.010	0.0003 ± 0.001	0.741 ± 0.243	0.0002 ± 0.0005	0.0004 ± 0.001
SlideVQA	Human sketches	0.747 ± 0.136	0.659 ± 0.226	0.238 ± 0.136	0.006 ± 0.010	0.787 ± 0.248	0.006 ± 0.010	0.236 ± 0.139
	Synthetic sketches	0.752 ± 0.132	0.637 ± 0.237	0.240 ± 0.134	0.008 ± 0.013	0.755 ± 0.271	0.006 ± 0.010	0.235 ± 0.138

Table 1. Comparison between Synthetic and Human Collected Sketches.

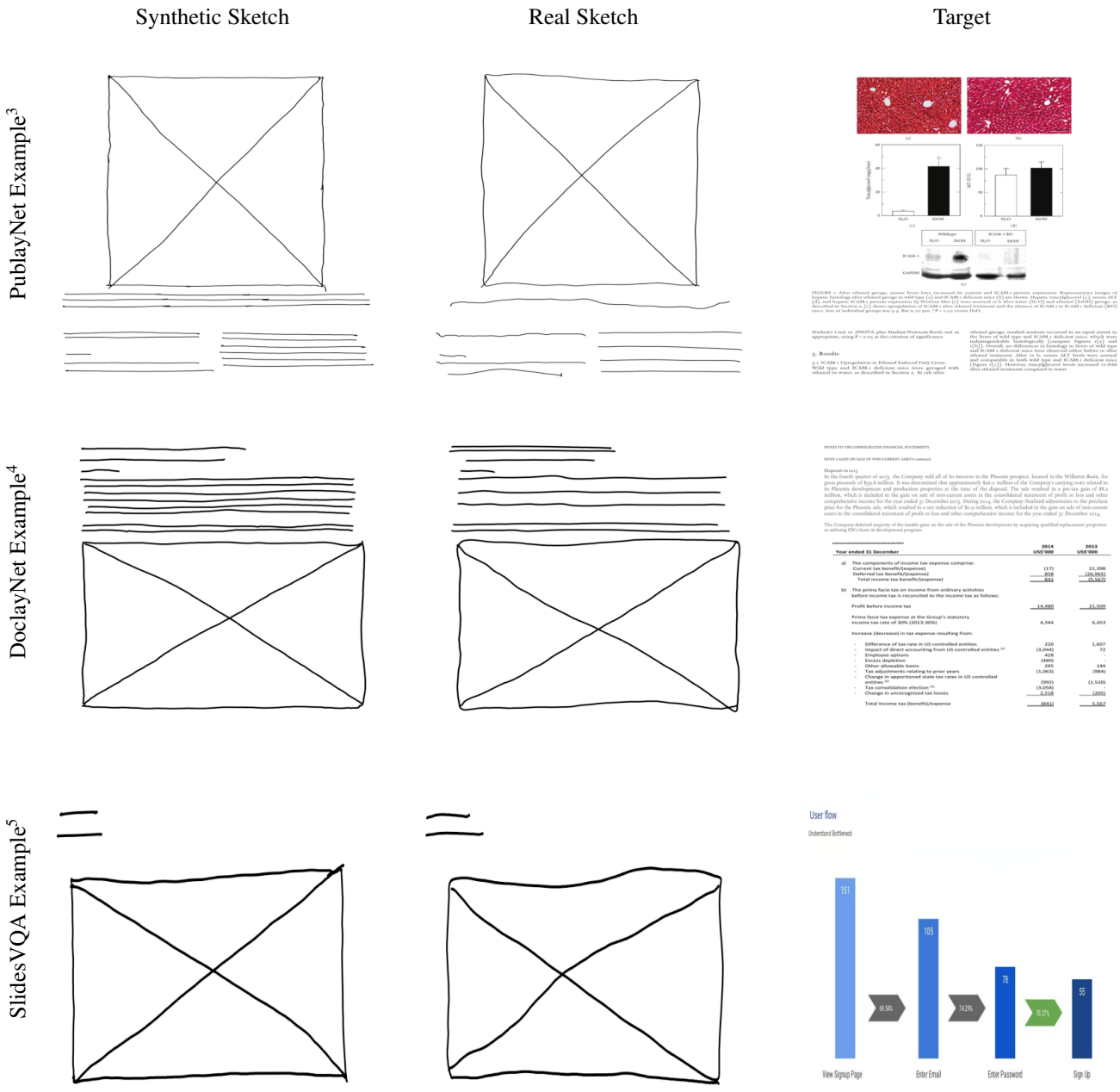

Figure 5. Some example layouts with corresponding synthetic and human collected sketches.

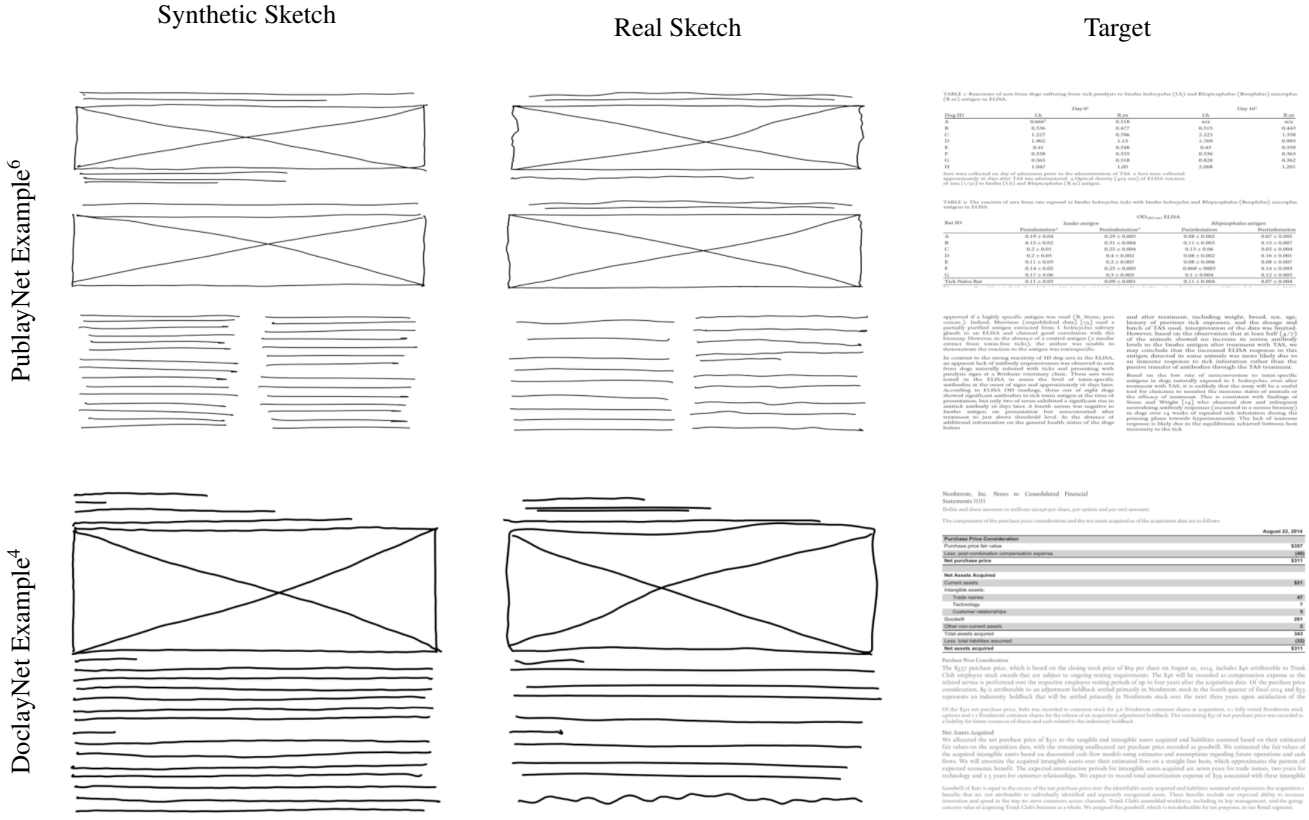

Figure 6. More example layouts with corresponding synthetic and human collected sketches.

7. Qualitative Results

Qualitative results of our method can be seen on Figure 7, 8 and 9 where the assets are shown as boxes with different colors specifying different assets. It can be seen that our method can generate layouts which are more accurate both in terms of the positioning and the ordering of the assets compared to LayoutPrompter(Gen-T, Gen-TS, Gen-R) and few-shot Gemini.

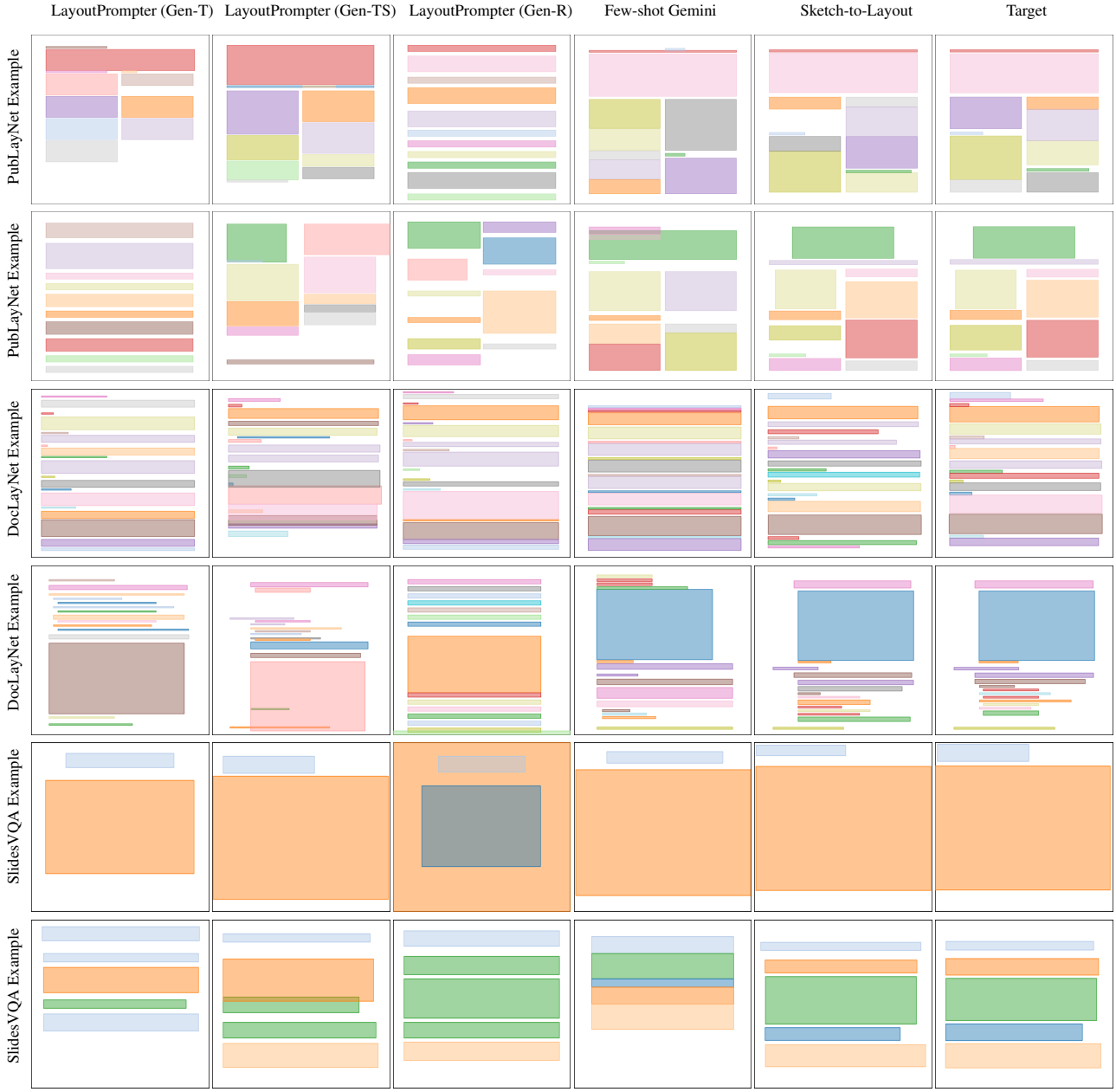

Figure 7. Examples of layouts generated by different methods and our model given the set of assets. Different assets are identified with different colors, showing the capability of different models to process asset content.

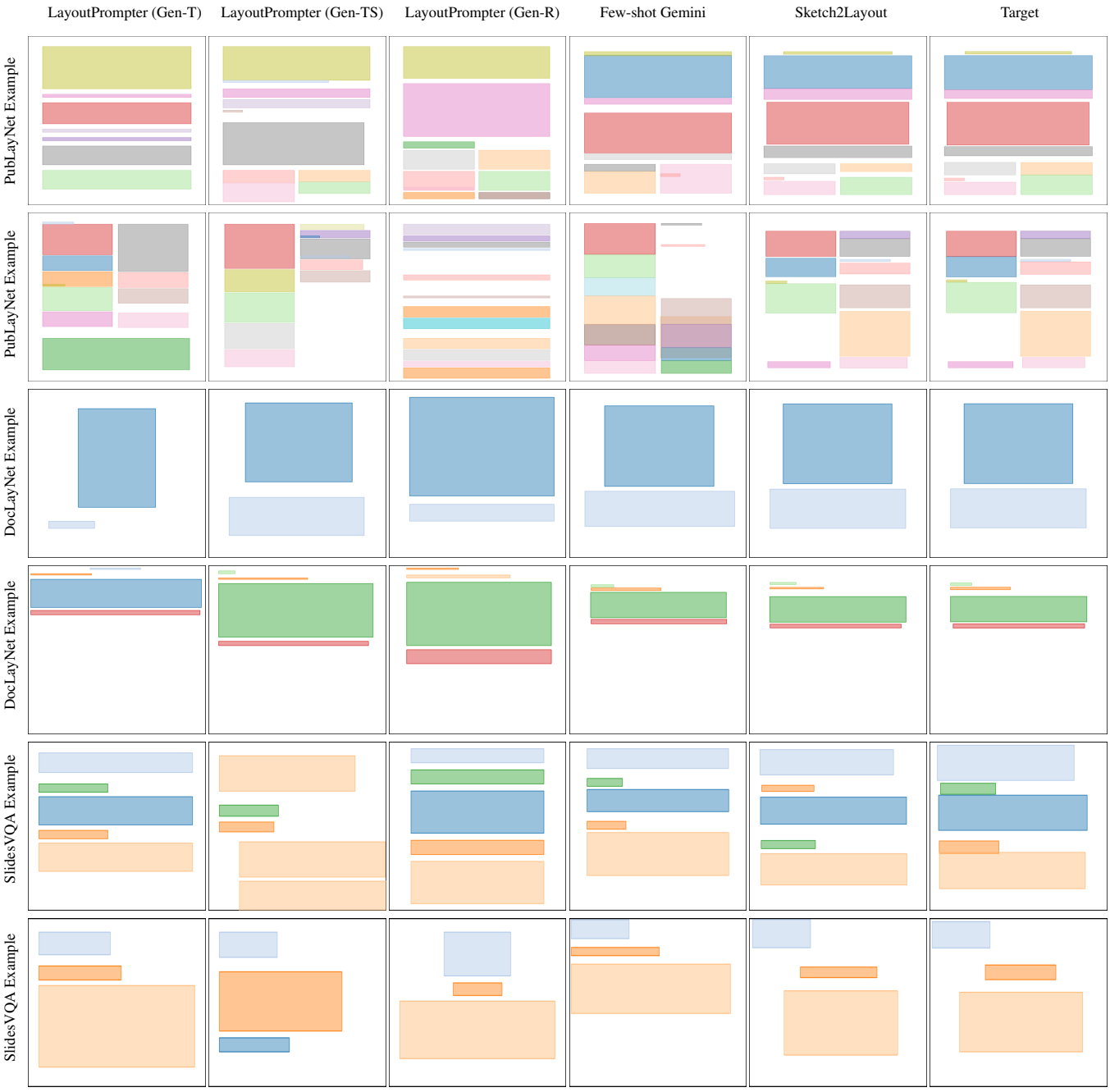

Figure 8. More examples of layouts generated by different methods and our model given the set of assets.

Figure 9. More examples of layouts generated by different methods and our model given the set of assets.

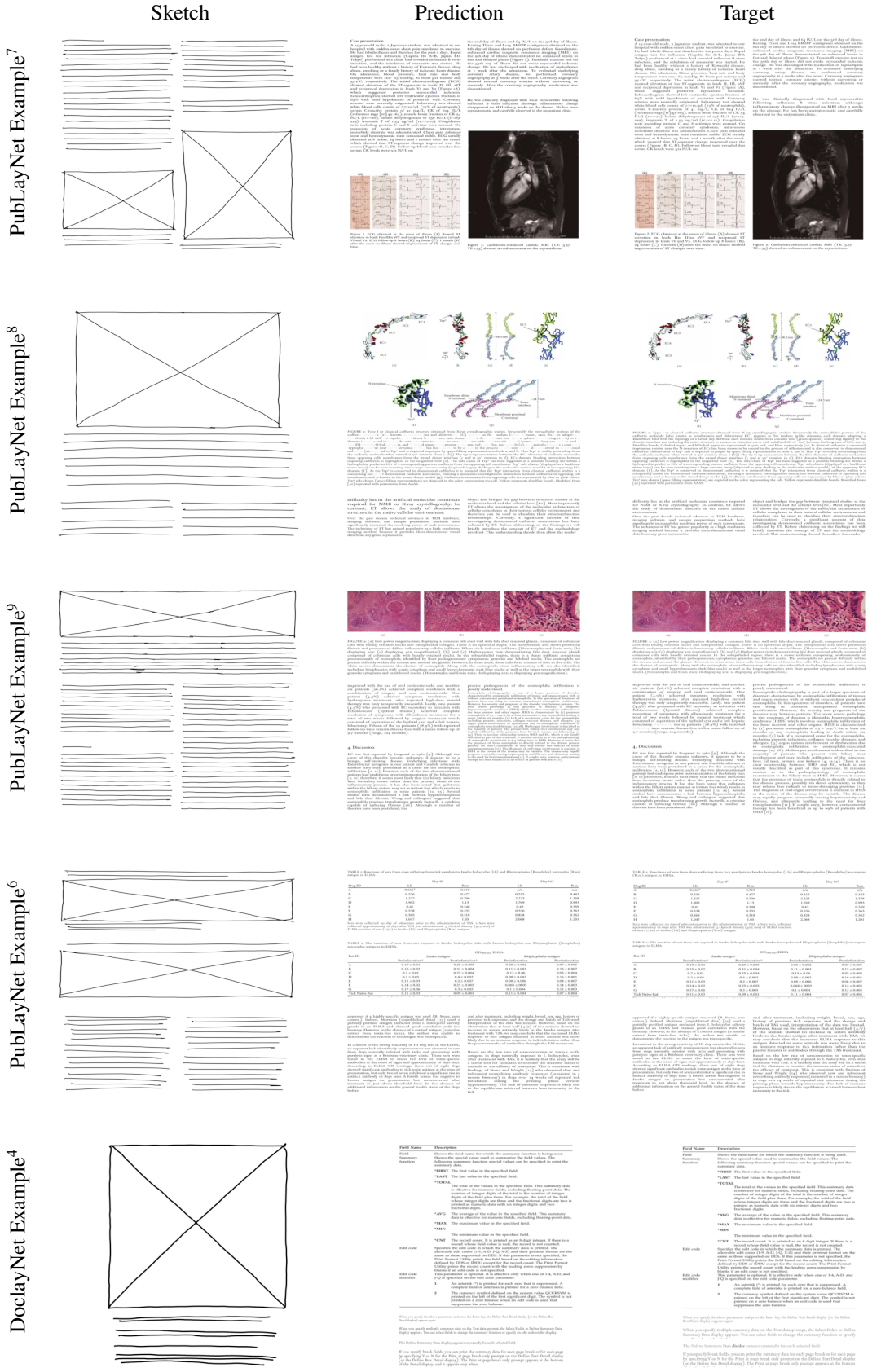

Figure 10. Sketches with corresponding predictions and the target layouts. Our method is able to generate layouts that conform to the sketch and have meaningful semantic order.

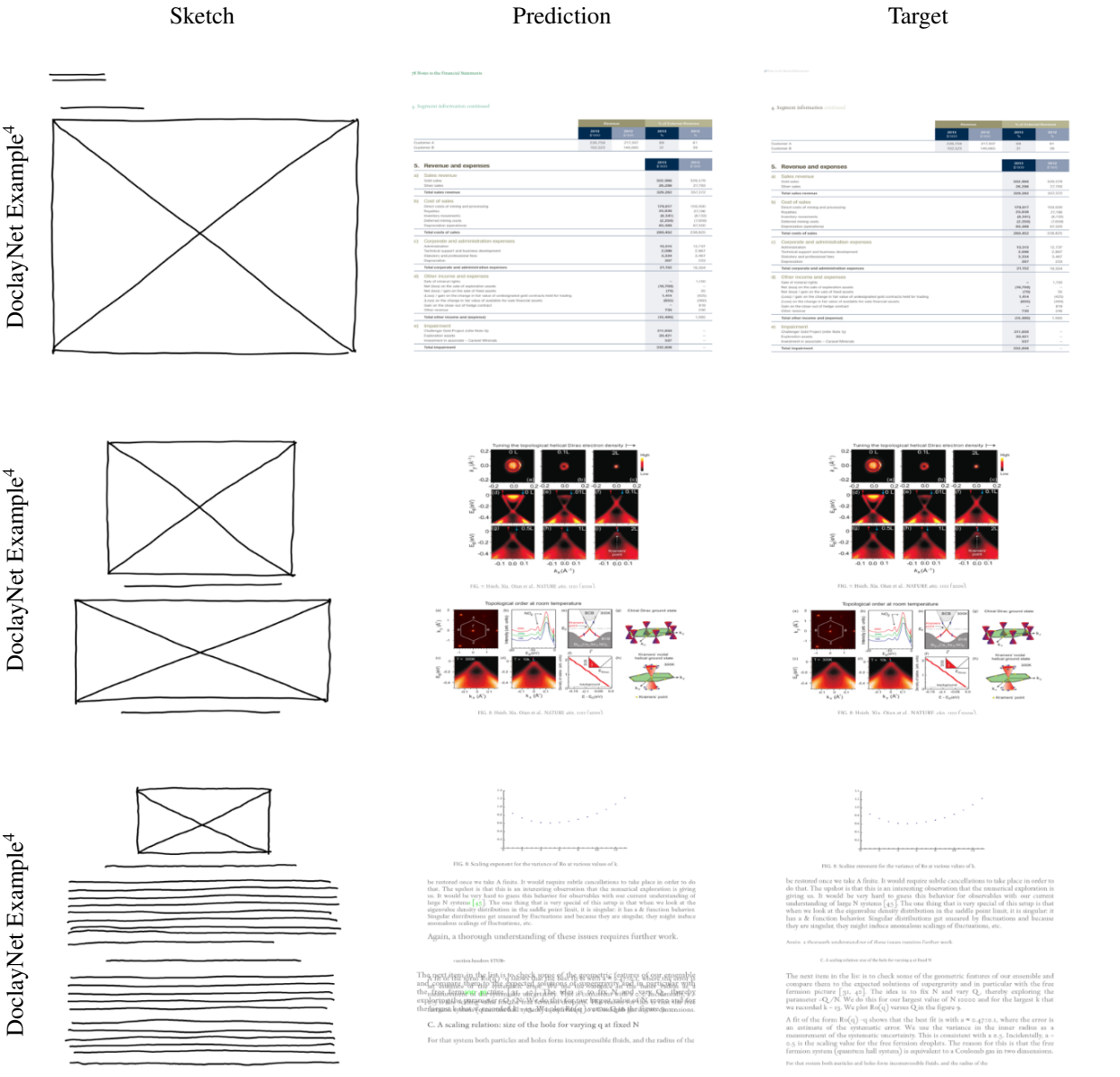

Figure 11. Sketches with corresponding predictions and the target layouts.

8. Legal Attributions

References

[1] Jiawei Lin, Jiaqi Guo, Shizhao Sun, Zijiang Yang, Jian-Guang Lou, and Dongmei Zhang. LayoutPrompter: Awaken the Design Ability of Large Language Models. *Advances in Neural Information Processing Systems*, 36, 2024. 1

[2] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Warm Restarts, 2017. 1

[3] Songrit Maneewongvatana and David M Mount. Analysis of Approximate Nearest Neighbor Searching with Clustered Point Sets. *arXiv preprint cs/9901013*, 1999. 1

[2]Target image licensed under CC-BY license available at http://creativecommons.org/licenses/by/2.0. Handschel J, Naujoks C, Depprich R, Lammers L, Kübler N, Meyer U, Wiesmann HP. Embryonic stem cells in scaffold-free three-dimensional cell culture: osteogenic differentiation and bone generation. Head Face Med. 2011 Jul 14;7:12. doi: 10.1186/1746-160X-7-12. PMID: 21752302; PMCID: PMC3143924 [3]Target image licensed under CC-BY license available at http://creativecommons.org/licenses/by/4.0. Theruvath TP, Ramshesh VK, Zhong Z, Currin RT, Karrasch T, Lemasters JJ. Icam-1 upregulation in ethanol-induced Fatty murine livers promotes injury and sinusoidal leukocyte adherence after transplantation. HPB Surg. 2012;2012:480893. doi: 10.1155/2012/480893. Epub 2012 Jun 18. PMID: 22778492; PMCID: PMC3385666 [4]DocLayNet examples are licensed under CDLA - Permissive, Version 1.0, available at https://cdla.dev/permissive-1-0/. DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis Pfitzmann, Birgit and Auer, Christoph and Dolfi, Michele and Nassar, Ahmed S and Staar, Peter W J [5]SlideVQA examples are licensed under CC-BY available at https://creativecommons.org/licenses/by/4.0/. SlideVQA: A Dataset for Document Visual Question Answering on Multiple Images, Tanaka, Ryota, Nishida, Kyosuke, Hasegawa, Taku, Saito, Itsumi, Saito, Kuniko, 2023/01/12 [6]Target image licensed under CC-BY license available at http://creativecommons.org/licenses/by/4.0. Hall–Mendelin S, O'Donoghue P, Atwell RB, Lee R, Hall RA. An ELISA to Detect Serum Antibodies to the Salivary Gland Toxin of Ixodes holocyclus Neumann in Dogs and Rodents. J Parasitol Res. 2011;2011:283416. doi: 10.1155/2011/283416. Epub 2011 May 18. PMID: 21687655; PMCID: PMC3112514. [7]Target image licensed under CC-BY license available at http://creativecommons.org/licenses/by/3.0. Muneuchi J, Kanaya Y, Takimoto T, Hoshina T, Kusuhara K, Hara T. Myocarditis mimicking acute coronary syndrome following influenza B virus infection: a case report. Cases J. 2009 Jun 25;2:6809. doi: 10.4076/1757-1626-2-6809. PMID: 19829864; PMCID: PMC2740055. [8]Target image licensed under CC-BY license available at https://creativecommons.org/licenses/by/4.0/. Owen GR, Stokes DL. Exploring the Nature of Desmosomal Cadherin Associations in 3D. Dermatol Res Pract. 2010;2010:930401. doi: 10.1155/2010/930401. Epub 2010 Jun 21. PMID: 20672011; PMCID: PMC2905946. [9]Target image licensed under CC-BY license available at http://creativecommons.org/licenses/by/4.0. Nashed C, Sakpal SV, Shusharina V, Chamberlain RS. Eosinophilic cholangitis and cholangiopathy: a sheep in wolves clothing. HPB Surg. 2010;2010:906496. doi: 10.1155/2010/906496. Epub 2010 Nov 7. PMID: 21076681; PMCID: PMC2976516.