A. Appendix Overview

The appendix has the following contents:

- Vision demonstrations of UDOP localizing answers in documents, the effectiveness of the cross attention with character embeddings in vision generation, and more neural editing examples Appendix B.
- Downstream evaluation datasets details in Appendix C.
- UDOP-Dual performance in Appendix D.
- More details for pretraining and evaluation datasets, and finetuning experiment set up in Appendix E.
- Few shot learning in Appendix F.
- Effectiveness of the Vision Modality in Appendix G.
- Additional Supervised Training Stage in Appendix H.
- Experiment results of curriculum learning in Appendix I.
- Performance variance of UDOP in Appendix J.
- Discussion of limitations and societal impacts in Appendix K.

B. Visualization Analysis

Creative Image Generation. UDOP achieves controllable high-quality document generation and editing as described in Section 6.1. We show additional examples here in Fig. 7. Our model can edit and add to the document image content with customized contents. Note that even if the document content is vertical (the first subfigure of Fig. 7), UDOP can still achieve high generation quality.

Layout Customization. UDOP can perform controllable high-quality document layout edits. We show examples in Figure 6, where our model can edit the layout of the document by regenerating the document from scratch. This is done by keeping only a few image patch as prompt, change the bounding boxes of the content, and then regenerate the document image with the new layout.

Answer Localization for Document QA. UDOP can perform question answering while predicting the location of the answer. We show examples on VisualMRC in Figure 8 and our model can answer the questions regarding the document correctly while locating the area of interest.

C. Downstream Evaluation Datasets

FUNSD (Form Understanding in Noisy Scanned Documents [18]) has 149 and 50 samples for train and test. We evaluate on the entity recognition task: predicting the entity, "question", "answer", "header", or "other", for the text token. The task format is, suppose we have the title, "The Title", and its entity "[I-Header]", then the encoder input is "The Title" and the generation target is "The Title [I-Header]". The metric is F1 scores.

CORD (Consolidated Receipt Dataset for Post-OCR Parsing) [34] is a key information extraction dataset with 30 labels under 4 categories such as "total" or "subtotal". It has 1,000 receipt samples. The train, validation, and test splits contain 800, 100, and 100 samples respectively. The metric is F1 and the task format is the same as FUNSD.

RVL-CDIP is the document classification dataset that we have discussed previously. It has 320k/40k/40k images for training/validation/test. The metric is classification accuracy.

DUE-Benchmark contains 7 datasets and 3 domains, including document question answering (DocVQA [33], InfographicsVQA [32]), key information extraction (KLC [41], PWC [19], DeepForm [43]), and Table QA/NLI (WTQ [35], TabFact [5]). Task prompt formats can be found in Section 4.2 and details of datasets can be found in the appendix.

D. UDOP-Dual Performance

We list the performance of UDOP-Dual on FUNSD, CORD, and RVL-CDIP in Table 9.

E. Supervised Pretraining Tasks

In this section, we list more details about the supervised datasets in pretraining and evaluations.

E.1. Classification

RVL-CDIP [13] contains 16 document categories, such as “invoice”, “scientific publication” and “form”. The dataset has 320k training, 40k validation and 40k test images.

E.2. Layout Analysis

PubLayNet [57] is a layout analysis dataset created from medical publications. It contains over 360k document images and labeled with typical document layout elements such as titles, paragraphs, etc.

E.3. Information Extraction

DocBank [28] is a richly-annotated large-scale IE dataset. It consists of 500K document pages, where 400K for training, 50K for validation and 50K for testing. It has 12 semantic structure labels like abstract, title, and author. Each token has corresponding bounding box and semantic structure label.

Kleister Charity [41] is an IE dataset with complex invoice page layout and has 21.6k entities and 2.7k document images from UK Charity Commission. Its entities for extraction include invoice date, invoice number, net amount, vendor name, etc.

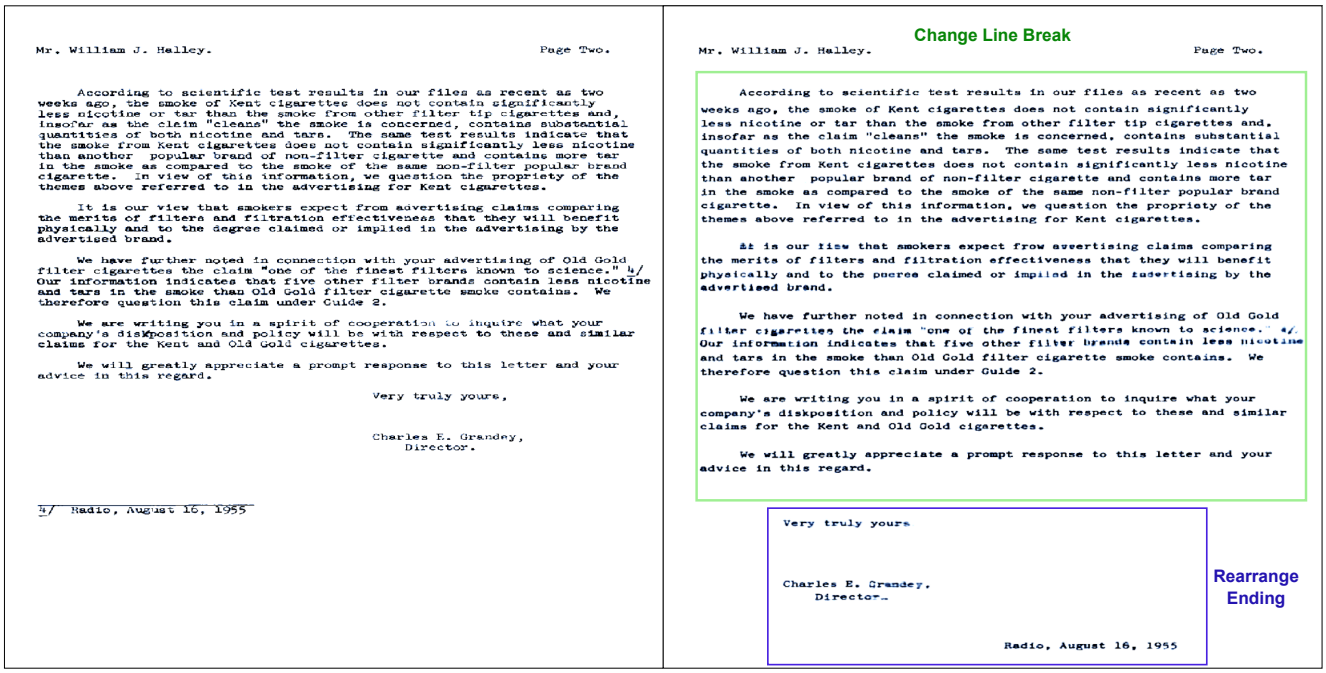

Figure 6. Document generation with customized layout (right). Left is the original document. We change the layout of the document text including line breaks change and text rearrangement. All edits are done with one model run.

Table 6. Comparison of different image size in curriculum learning on the DUE-Benchmark. Modality T, L, V denote text, layout, or vision.

Model	Modality	Question Answering		Information Extraction			Table QA/NLI		Avg.
		DocVQA	InfoVQA	KLC	PWC	DeepForm	WTQ	TabFact	
UDOP (224)	V+T+L	84.4	46.1	82.1	26.7	83.6	46.1	78.2	63.9
UDOP (512)	V+T+L	84.5	47.3	82.0	27.1	84.7	46.2	78.3	64.3
UDOP (1024)	V+T+L	**84.7**	**47.4**	**82.8**	**28.9**	**85.5**	**47.2**	78.9	**65.1**

PWC [19] is an IE dataset which has 2,291 leaderboards, where the data is collected from the Papers with Code labelling interface. It asks information like task, dataset, metric, etc. Different from original implementation, DUE-Benchmark provides complete papers as input instead of tables.

DeepForm [43] is an IE dataset collected from political television ads in US elections and has 20k receipts and over 100k document images. This task is to extract entities like advertiser name, contract number, amount paid, etc.

E.4. Question Answering

WebSRC [3] stands for Web-based Structural Reading Comprehension. It consists of 0.44M questions collected from 6.5K web pages with corresponding HTML, screenshots and metadata. The answer is either the text span of context or yes/no.

VisualMRC [45] stands for visual machine reading comprehension. It consists of 10,197 images 30,562 abstractive questions-answers.

DocVQA [33] is a QA dataset for excerpts from industry documents and has 50k questions on 12k document images. It asks questions on topics like text content, non-textual elements like marks or diagrams, layout, style, etc.

InfographicsVQA [32] is a QA dataset with a focus on infographic images and has 30K questions on 5.3k document images. It requires reasoning on text content, images, data visualizations, layout, etc.

WTQ [35] is a table-based QA dataset on HTML tables collected from Wikipedia. It has 2.1k tables and 22k questions hand crafted by humans and cover a wide range of topics like table lookup, superlatives, arithmetic operations, etc.

E.5. Document NLI

TabFact [5] is an open-domain table-based NLI task and has 16k Wikipedia tables for 118k statements by human annotations.

Table 7. Performance with standard deviations on on the DUE-Benchmark. Modality T, L, V denote text, layout, or vision.

Model	Modality	Question Answering		Information Extraction			Table QA/NLI		Avg.
		DocVQA	InfoVQA	KLC	PWC	DeepForm	WTQ	TabFact	
Donut	V	72.1	-	-	-	-	-	-	-
$\text{BERT}_{\text{large}}$ [9]	T	67.5	-	-	-	-	-	-	-
T5_{large} [39]	T	70.4	36.7	74.3	25.3	74.4	33.3	58.9	50.7
T5_{large}+U [36]	T	76.3	37.1	76.0	27.6	82.9	38.1	76.0	56.5
T5_{large}+2D [36]	T+L	69.8	39.2	72.6	25.7	74.0	30.8	58.0	50.4
T5_{large}+2D+U [36]	T+L	81.0	46.1	75.9	26.8	83.3	43.3	78.6	59.8
LAMBERT [10]	T+L	-	-	81.3	-	-	-	-	-
$\text{StructuralLM}_{\text{large}}$ [26]	T+L	83.9	-	-	-	-	-	-	-
$\text{LayoutLMv2}_{\text{large}}$ [55]	V+T+L	78.8	-	-	-	-	-	-	-
$\text{LayoutLMv3}_{\text{large}}$ [16]	V+T+L	83.4	45.1	77.1	26.9	84.0	45.7	78.1	62.9
UDOP-Dual	V+T+L	84.4±0.1	47.1±0.2	81.9±0.4	28.7±0.5	85.2±0.2	46.7±0.4	**79.5**±0.3	64.7±0.3
UDOP	V+T+L	**84.7**±0.2	**47.4**±0.2	**82.8**±0.3	**28.9**±0.4	**85.5**±0.2	**47.2**±0.2	78.9±0.1	**65.1**±0.2

Table 8. Performance with standard deviations on FUNSD, CORD, and RVL-CDIP datasets.

Model	Modality	Info Ext.		Classification
		FUNSD	CORD	RVL-CDIP
Donut	V	-	91.6	95.3
$\text{BERT}_{\text{large}}$	T	65.63	90.25	89.92
$\text{BROS}_{\text{large}}$ [15]	T+L	84.52	97.40	-
$\text{StructuralLM}_{\text{large}}$	T+L	85.14	-	96.08
LiLT [48]	T+L	88.41	96.07	95.68
FormNet [24]	T+L	84.69	97.28	-
$\text{LayoutLM}_{\text{large}}$	T+L	77.89	-	91.90
SelfDoc	V+T+L	83.36	-	92.81
UDoc	V+T+L	87.93	98.94	95.05
$\text{DocFormer}_{\text{large}}$ [1]	V+T+L	84.55	96.99	95.50
$\text{TILT}_{\text{large}}$	V+T+L	-	96.33	95.52
$\text{LayoutLMv2}_{\text{large}}$	V+T+L	84.20	96.01	95.64
$\text{LayoutLMv3}_{\text{large}}$	V+T+L	92.08	97.46	95.93
UDOP-Dual	V+T+L	91.20±0.21	97.64±0.12	96.22±0.27
UDOP	V+T+L	91.62±0.34	97.58±0.15	96.00±0.26

E.6. Finetuning Experiment Setting

For all DUE-Benchmark finetuning experiments, we use Adam [23] optimizer with learning rate 5e-5, 1000 warmup steps, batch size 16, weight decay of 1e-2, $\beta_1 = 0.9$, and $\beta_2 = 0.98$. For FUNSD and CORD, we use learning rate 3e-4 and for RVL-CDIP, we use learning rate 1e-3 both with 1000 warmup steps, batch size 16, weight decay of 1e-2, $\beta_1 = 0.9$, and $\beta_2 = 0.98$.

F. Few-shot Learning

UDOP has few-shot ability on unseen datasets. See Table 10 for more details on few-shot performance on FUNSD and Tobacco-3482, which are not included in the pretraining. FUNSD is introduced in Section 5.2 and has 199 samples. The Tobacco-3482 dataset has document images with 10 classes such as email, letter, form, etc. The dataset has 3482 images including 2.7k training samples and 0.7k testing samples.

G. Effectiveness of the Vision Modality

In the field of Document AI, the effectiveness of the vision modality, i.e., document images, is unclear. We explore this by removing the visual embedding from the model input, with results shown in Table 11. It shows that the vision modality is more prominent on visually-rich tasks, e.g., InfographicsVQA, compared with text-dominant data such as DocVQA.

Table 9. Performance of UDOP-Dual on FUNSD, CORD, and RVL-CDIP.

Model	Modality	Info Ext.		Classification
		FUNSD	CORD	RVL-CDIP
Donut [21]	V	-	91.6	95.3
$\text{BERT}_{\text{large}}$ [9]	T	65.63	90.25	89.92
$\text{BROS}_{\text{large}}$ [15]	T+L	84.52	97.40	-
$\text{StructuralLM}_{\text{large}}$ [26]	T+L	85.14	-	96.08
LiLT [48]	T+L	88.41	96.07	95.68
FormNet [24]	T+L	84.69	97.28	-
$\text{LayoutLM}_{\text{large}}$ [53]	T+L	77.89	-	91.90
SelfDoc [29]	V+T+L	83.36	-	92.81
UniDoc [11]	V+T+L	87.93	96.86	95.05
$\text{DocFormer}_{\text{large}}$ [1]	V+T+L	84.55	96.99	95.50
$\text{TILT}_{\text{large}}$ [36]	V+T+L	-	96.33	95.52
$\text{LayoutLMv2}_{\text{large}}$ [55]	V+T+L	84.20	96.01	95.64
$\text{LayoutLMv3}_{\text{large}}$ [16]	V+T+L	**92.08**	97.46	95.93
UDOP-Dual	V+T+L	91.20	97.64	**96.22**
UDOP	V+T+L	91.62	**97.58**	96.00

Table 10. Few-Shot Learning on FUNSD and Tobacco-3482.

Model	# Samples Per Class	FUNSD	Tobacco-3482
UDOP	All	91.6	96.0
UDOP	3	86.1	92.1
UDOP	1	82.4	87.6

Table 11. Effectiveness of the vision modality.

Model	DocVQA	InfoVQA
UDOP	**84.7**	**47.4**
UDOP w/o image input embeddings	84.4	45.0

H. Additional Supervised Training Stage

TILT [36] performs additional training on a wide range of QA datasets, such as reading comprehension dataset SQuAD [40], before the finetuning on DocVQA. This results in considerable performance improvement of the TILT model on DocVQA and InfographicsVQA. To have a fair comparison, we also finetune UDOP on the same set of datasets before testing on DocVQA or InfographicsVQA. As shown in Table 12, UDOP is further improved with this auxiliary training and outperforms TILT.

Table 12. Training UDOP on auxiliary QA datasets as in TILT. The performance of UDOP on DocVQA and InfographicsVQA is further improved (performance without the auxiliary training was not reported in the TILT paper).

Model	DocVQA	InfoVQA
$\text{TILT}_{\text{large}}$(w/ auxiliary training)	87.1	61.2
UDOP (w/o auxiliary training)	84.7	47.4
UDOP (w/ auxiliary training)	**87.8**	**63.0**

I. Curriculum Learning

In this section, we present the results of curriculum learning of input image resolution (224, 512, 1024) on the validations sets of evaluation benchmarks. As shown in Table 6, while the model already performs competitively well on 224 resolution, its performance further increases on 512 and 1024.

J. Performance Variance

For results in Table 2 and Table 3, we report their standard deviations as shown in Table 7 and Table 8. The deviations are computed from 5 runs with different seeds for parameter initialization.

K. Limitations and Societal Impact

UDOP can assist users with document analysis, understanding and information extraction. This automatic processing technology will make the document processing workflow more efficient and potential more accurate. It is also worth noting that, similar to all AI generation technology, the document generation capacity of UDOP can be potentially abused for malicious document counterfeit, e.g., signature forgery, tampering monetary amount in checks, fake medical/financial records generation, etc. To avoid abuse, for model release we plan to open source the vision generation model only with limited access, e.g., through an API. Documents submitted by users that are classified as sensitive (the classifier can be a finetuned UDOP model), such as checks and personal ID, will be denied.

Applying UDOP on non-English data, especially those with non-Latin writing systems, may require further modifications to the model. For example, in Sec. 4.1, the vision decoder cross-attends with character embeddings. Then for non-English data, we need to include more character embeddings to attend with.

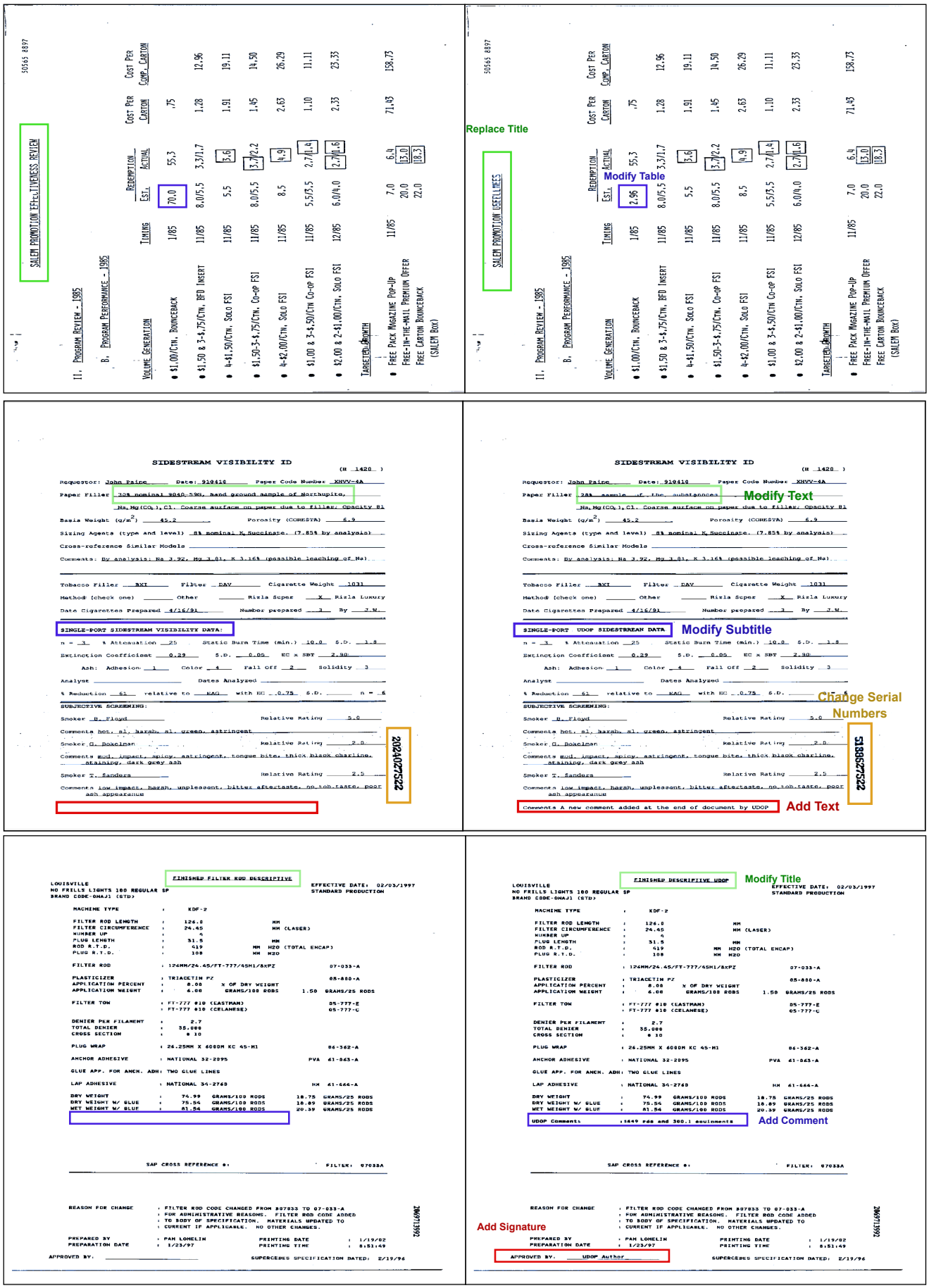

Figure 7. Document generation with customized content (right). Left is the original document. We show different document edits within the same figure including title replacement, text addition, text replacement, and tilted text replacement. All edits are done with one model run.

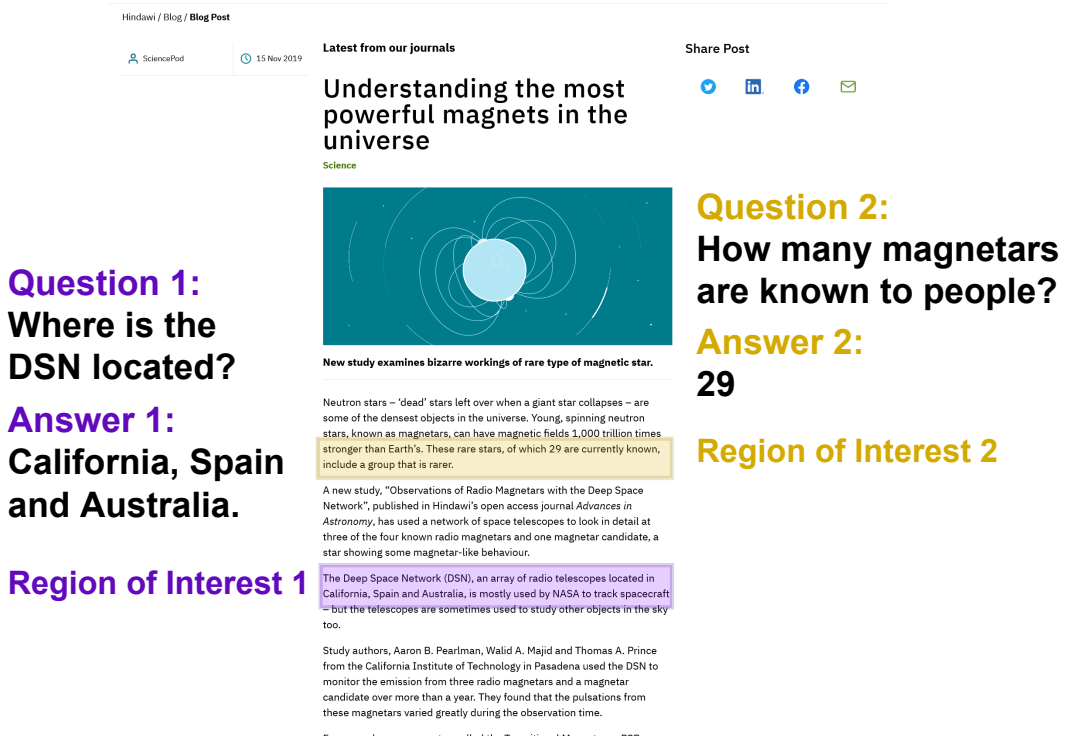

Figure 8. Document QA and answer localization with UDOP on VisualMRC dataset. As shown, besides generating the answer, UDOP can predict the region of interest (RoI) that answer is located in by generating the layout tokens. Note that the the labeled RoI VisualMRC dataset is at paragraph level.