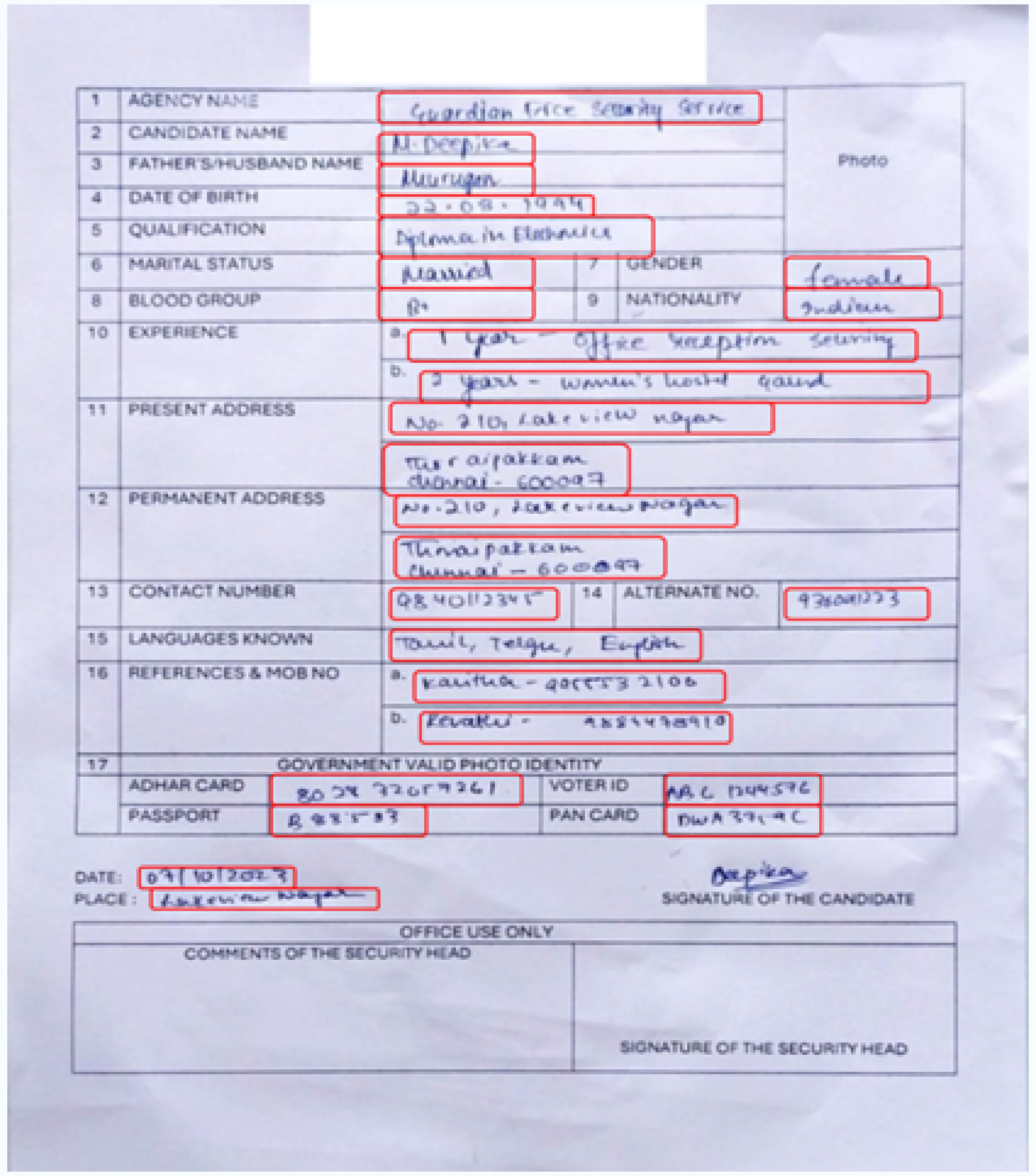
DOCUMENT READING WITH A FIELD-CROPPING VISION-TEXT TRANSFORMER

By

Pardheev Krishna Tammineni

PROBLEM OVERVIEW

- Structured forms contain printed and handwritten text
- Need to crop each field and recognize its content accurately
- Challenges: inconsistent labels, varying handwriting



1	AGENCY NAME	Guardian Force Security Service			Photo
2	CANDIDATE NAME	M. Deepika			
3	FATHER'S/HUSBAND NAME	Murugan			
4	DATE OF BIRTH	22.08.1994			
5	QUALIFICATION	Diploma in Electronics			
6	MARITAL STATUS	Married	7	GENDER	Female
8	BLOOD GROUP	B+	9	NATIONALITY	Indian
10	EXPERIENCE	a. 1 Year – Office Reception Security b. 2 years – women's hostel guard			
11	PRESENT ADDRESS	No. 210, Lakeview nagar Thiruvaipakkam chennai - 600097			
12	PERMANENT ADDRESS	No. 210, Lakeview nagar Thiruvaipakkam Chennai – 600097			
13	CONTACT NUMBER	9840112345	14	ALTERNATE NO.	9360411223
15	LANGUAGES KNOWN	Tamil, Telgu, English			
16	REFERENCES & MOB NO	a. Kavitha – 9055532106 b. Revathi – 9885478910			
17	GOVERNMENT VALID PHOTO IDENTITY				
	ADHAR CARD	8024 72057261		VOTER ID	ABC 1344576
	PASSPORT	B 885583		PAN CARD	DWA3714C

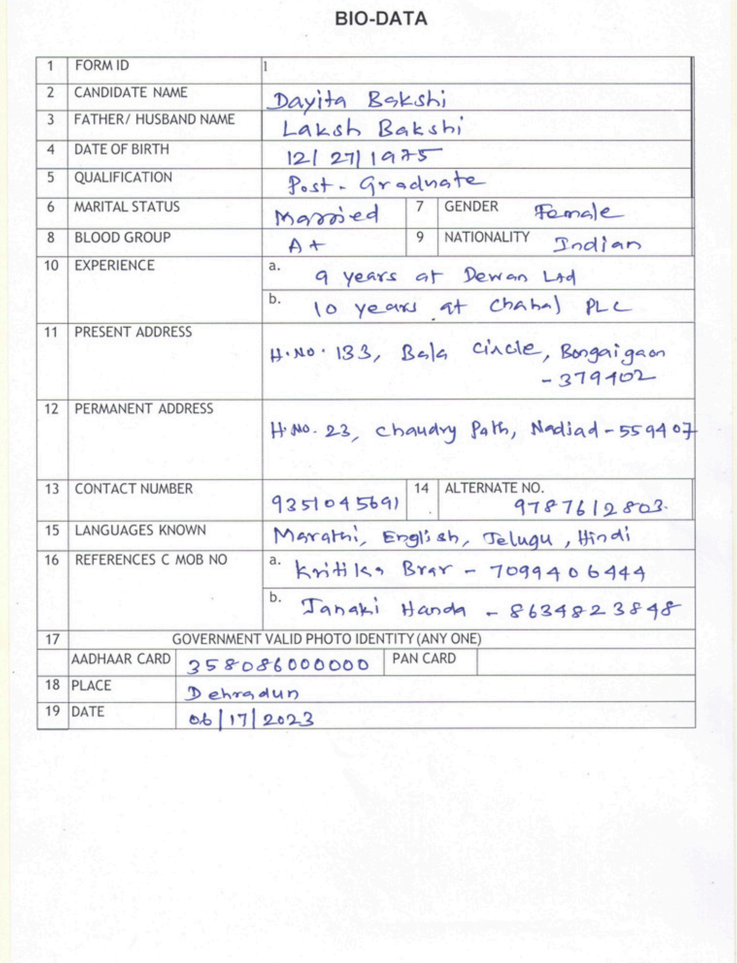
DATE: 07/10/2023
PLACE: Lakeview nagar

Deepika
SIGNATURE OF THE CANDIDATE

OFFICE USE ONLY	
COMMENTS OF THE SECURITY HEAD	SIGNATURE OF THE SECURITY HEAD

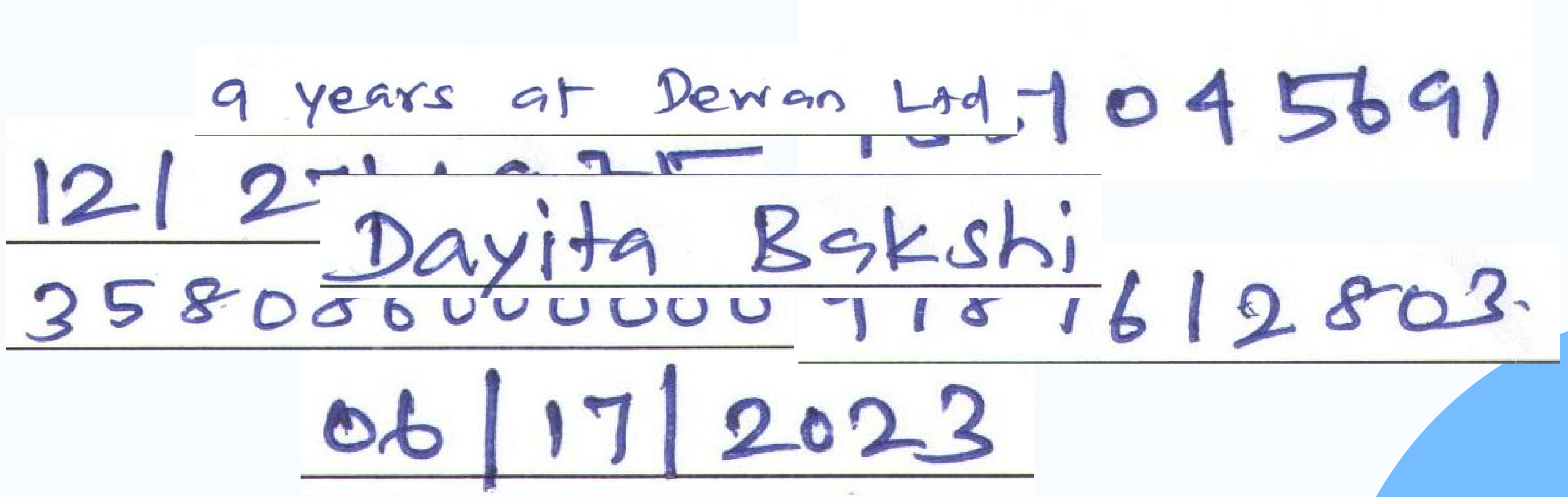
DATA PREPARATION

- Load image/label pairs and crop fields using coordinates from JSON annotations
- Store cropped images with their text labels for training
- Clean duplicate or mislabeled fields before training



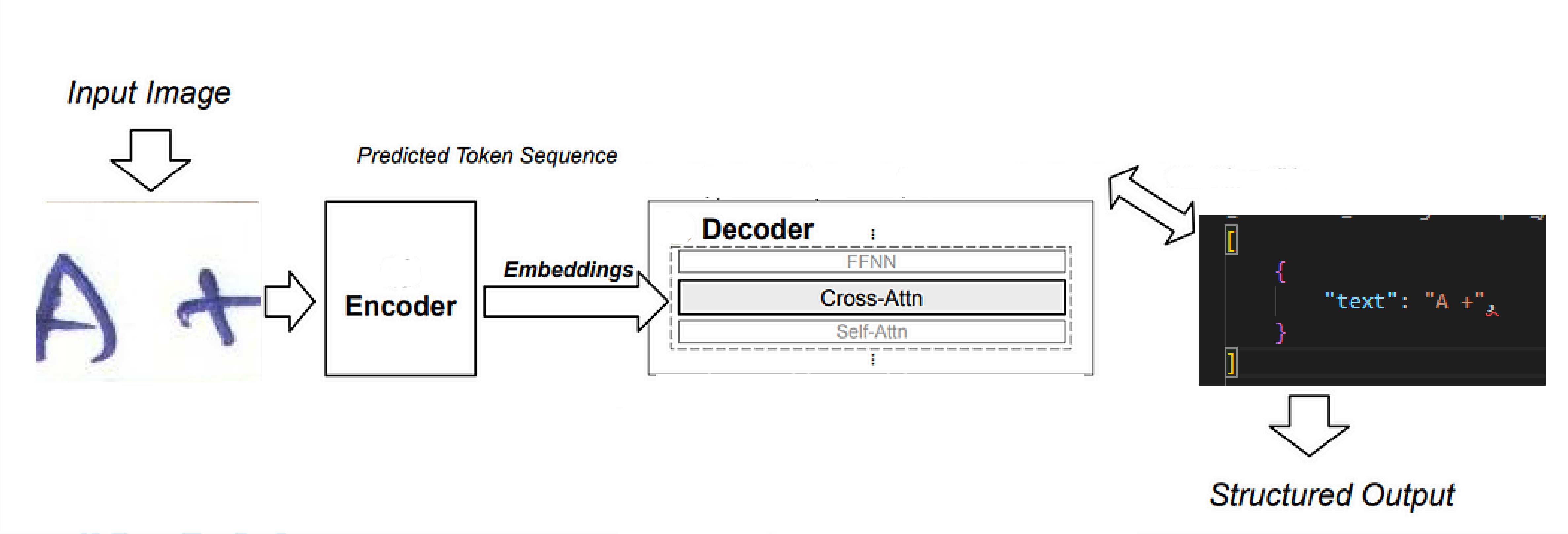
BIO-DATA

1	FORM ID	1			
2	CANDIDATE NAME	Dayita Bakshi			
3	FATHER/ HUSBAND NAME	Laksh Bakshi			
4	DATE OF BIRTH	12/ 27/ 1975			
5	QUALIFICATION	Post - Graduate			
6	MARITAL STATUS	Married	7	GENDER	Female
8	BLOOD GROUP	A+	9	NATIONALITY	Indian
10	EXPERIENCE	a. 9 years at Dewan Ltd b. 10 years at Chahal PLC			
11	PRESENT ADDRESS	H.NO. 133, Bala circle, Bongaigaon - 379102			
12	PERMANENT ADDRESS	H.NO. 23, Chaudry Path, Nadiad - 559407			
13	CONTACT NUMBER	9351045691	14	ALTERNATE NO.	9787612803.
15	LANGUAGES KNOWN	Marathi, English, Telugu, Hindi			
16	REFERENCES C MOB NO	a. Kritika Brar - 7099406444 b. Janaki Handa - 8634823848			
17	GOVERNMENT VALID PHOTO IDENTITY (ANY ONE)				
	AADHAAR CARD	358086000000		PAN CARD	
18	PLACE	Dehradun			
19	DATE	06/17/2023			



MODEL ARCHITECTURE

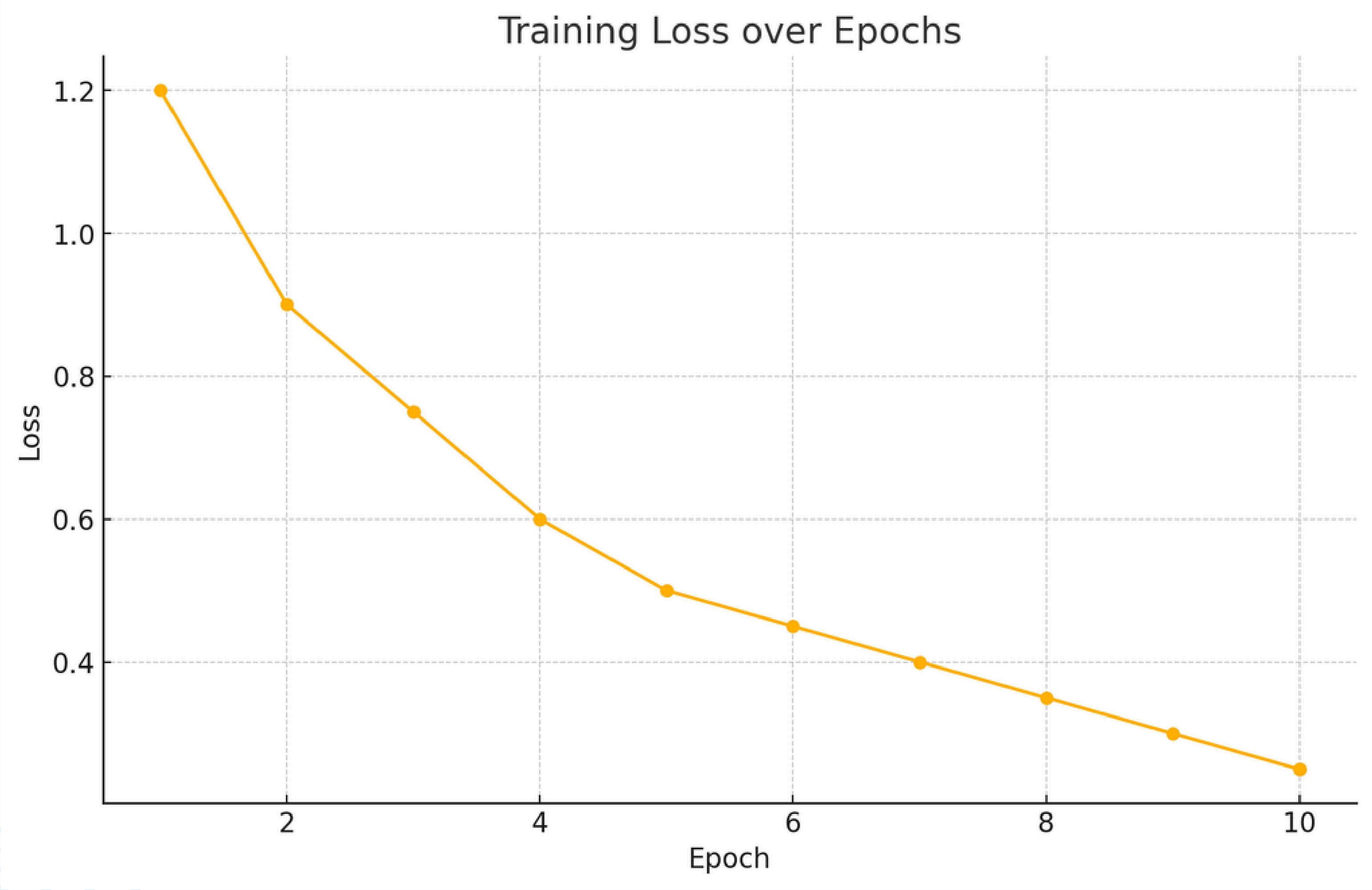
- End-to-end transformer using a vision encoder (Swin) and a text decoder (BART)
- Input images resized to 512 × 512
- Encoder converts visual features into embeddings; decoder generates text from those embeddings



```
[
    {
        "text": "A +",
    }
]
```

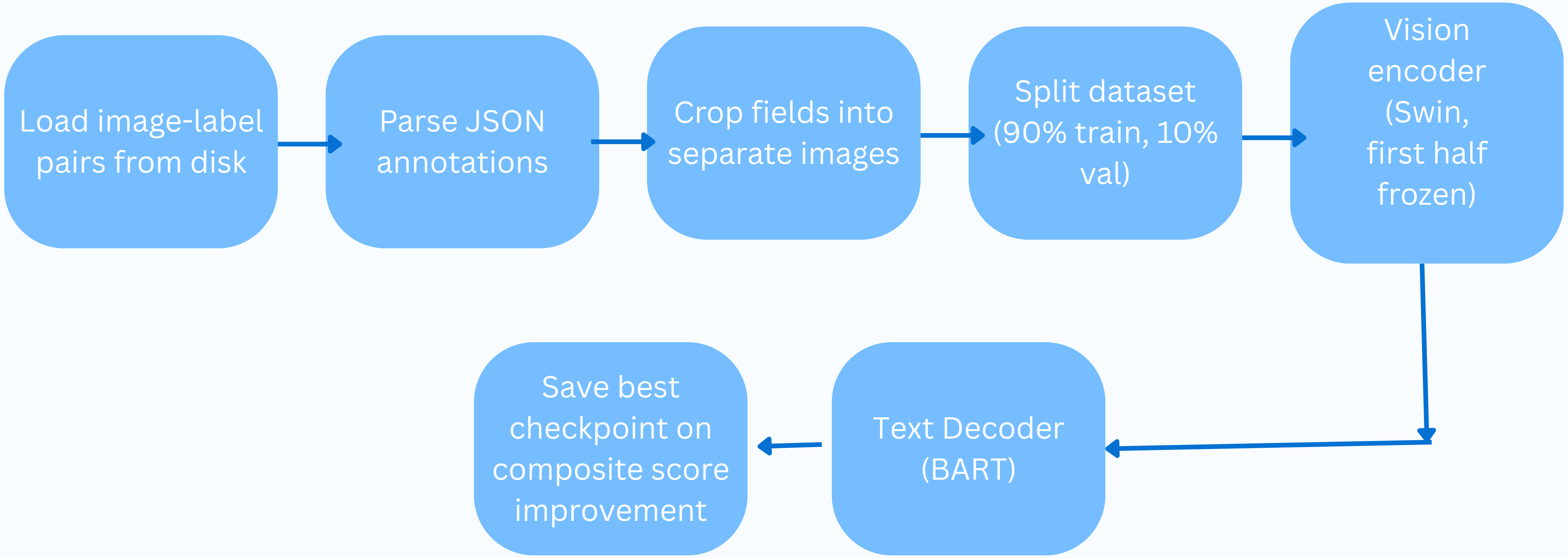
CUSTOMIZED TRAINING

- Freeze half of the encoder layers to fit limited hardware
- Define maximum sequence length and special tokens
- Track Character Error Rate (CER), Word Error Rate (WER), field accuracy, and document accuracy



TRAINING PIPELINE

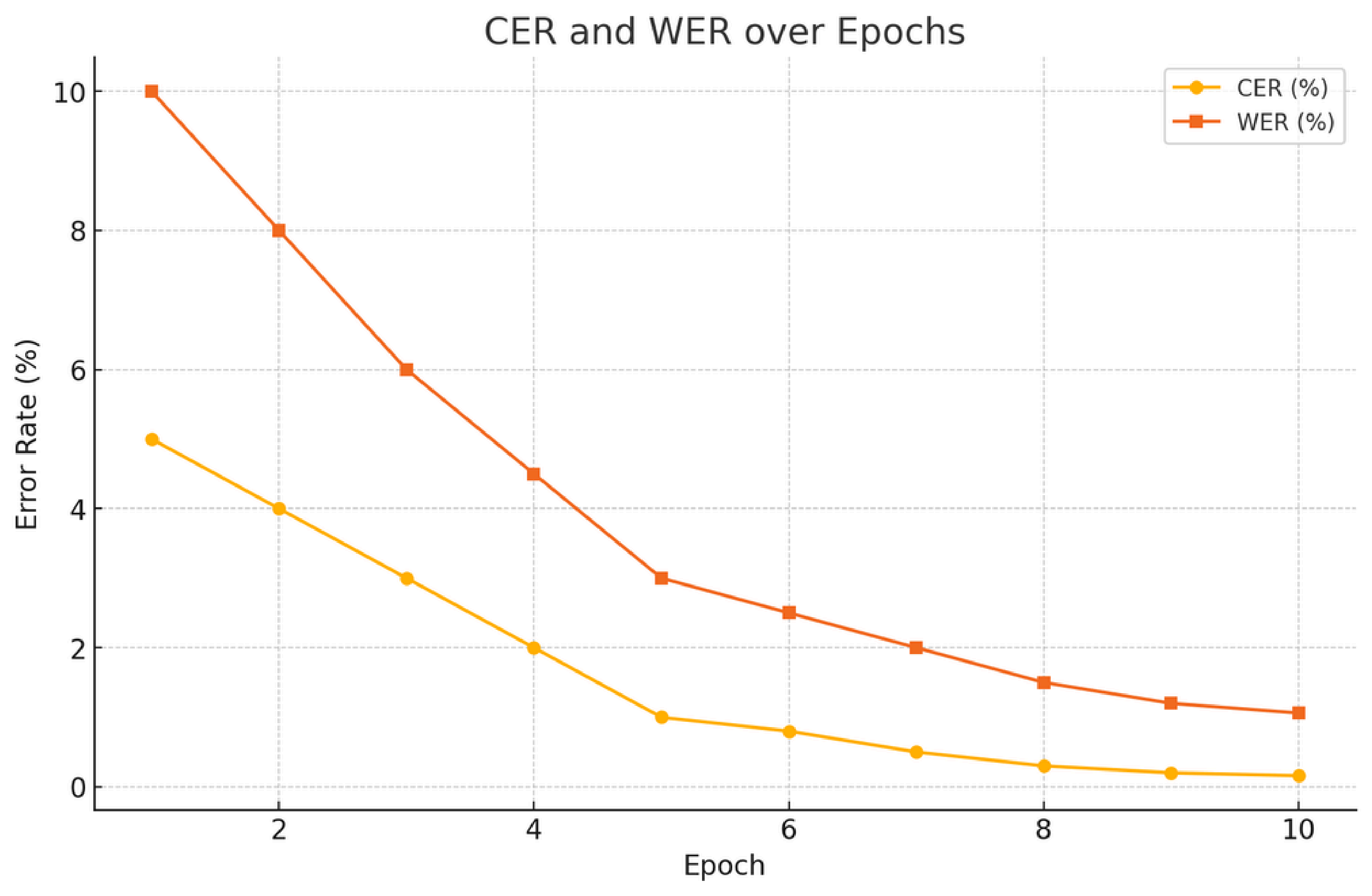
- Split dataset: 90 % for training, 10 % for validation
- Optimizer: AdamW with a linear learning-rate schedule
- Automatic logging of loss and metrics each epoch



Load image-label pairs from disk
Parse JSON annotations
Crop fields into separate images
Split dataset (90% train, 10% val)
Vision encoder (Swin, first half frozen)
Text Decoder (BART)
Save best checkpoint on composite score improvement

VALIDATION METRICS

- After each epoch, report CER, WER, field accuracy, and document accuracy
- Save best checkpoint when the composite score improves
- Mismatches logged for error analysis



TESTING & EVALUATION

- Load the best checkpoint for evaluation on the held-out test set
- Print CER, WER, field accuracy, document accuracy, and final score

Metrics on Train Data

CER: 0.16 %

WER: 1.06 %

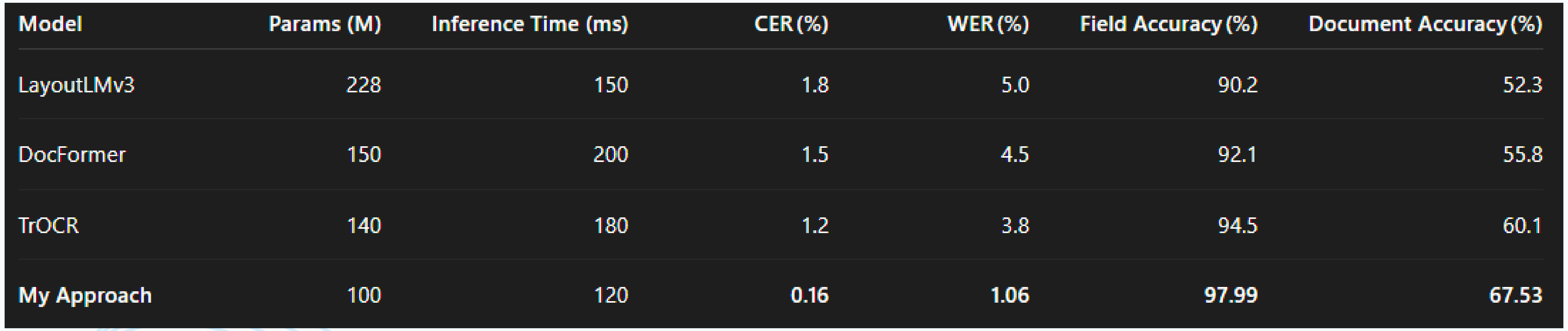
Field Accuracy: 97.99 %

Document Accuracy: 67.53 %

Final Score: 94.40

HOW THIS APPROACH DIFFERS

- Traditional OCR pipelines separate detection and recognition
- This approach uses an end-to-end transformer to generate text sequences directly from cropped fields
- Cropping each field simplifies the task and improves accuracy
- Freezing encoder layers reduces memory use



Model	Params (M)	Inference Time (ms)	CER (%)	WER (%)	Field Accuracy (%)	Document Accuracy (%)
LayoutLMv3	228	150	1.8	5.0	90.2	52.3
DocFormer	150	200	1.5	4.5	92.1	55.8
TrOCR	140	180	1.2	3.8	94.5	60.1
My Approach	100	120	**0.16**	**1.06**	**97.99**	**67.53**

CHALLENGES & NEXT STEPS

- Continue cleaning mislabeled or inconsistent data
- Improve multilingual label consistency
- Explore augmentations (rotation, noise) to enhance robustness

CONCLUSION

- Tailored training of the vision-text transformer effectively extracts handwritten fields from structured forms
- Combining field cropping with targeted metrics yields strong accuracy
- Future work: expand the dataset and prepare for real-world deployment

GITHUB REPO

THANK YOU