Scene Text Recognition

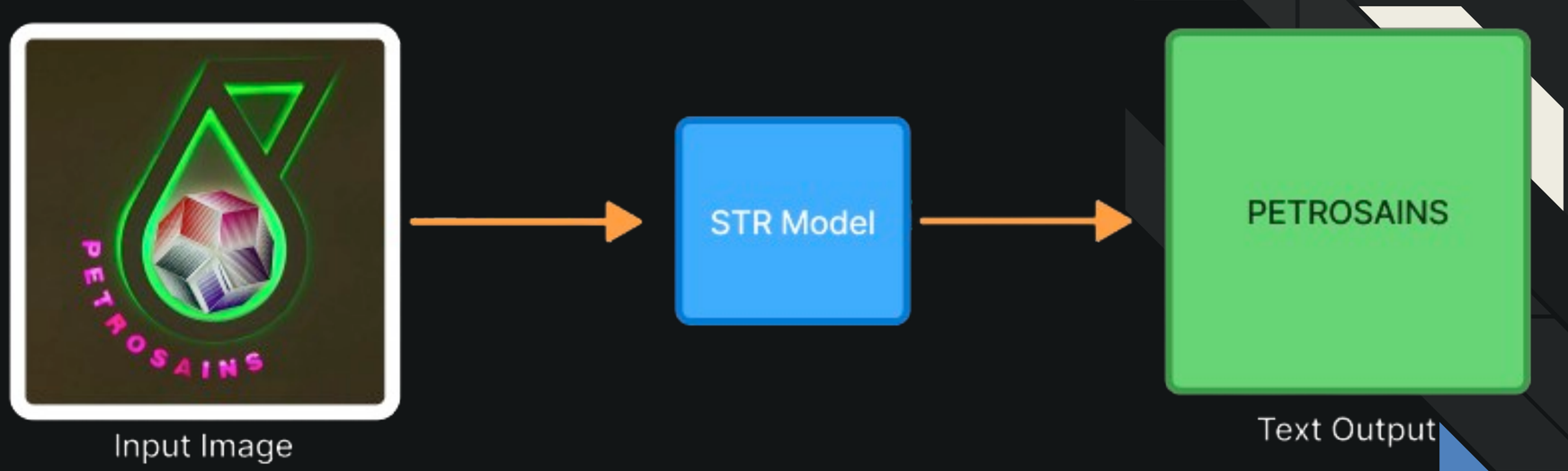
BETTER TEXT RECOGNITION
FOR COMPLEX SHAPES

Mentor: Anandita Jamwal

STR?

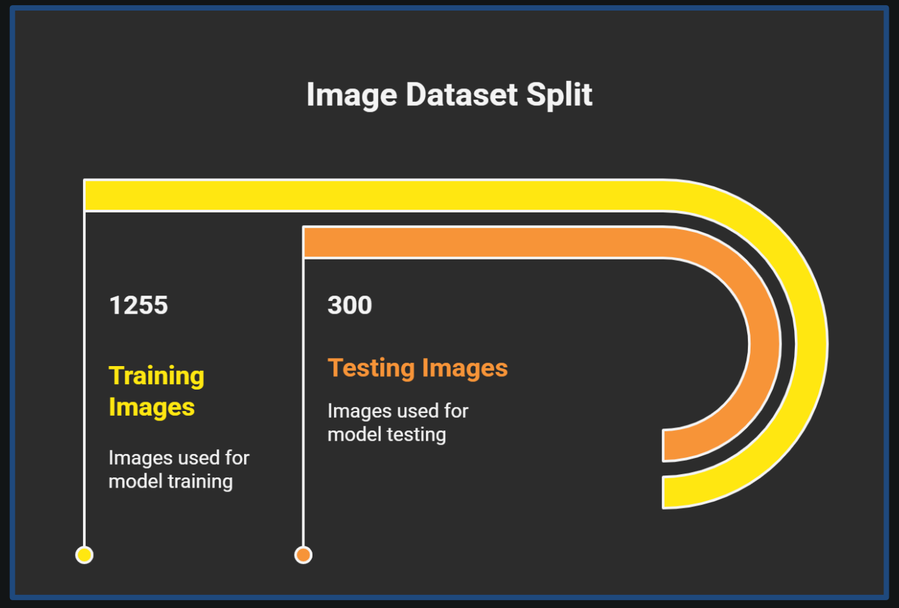
Scene Text Recognition (STR) is a crucial field within computer vision and natural language processing that aims to automatically detect and read text present in natural scene images. Unlike scanned documents, scene text appears in diverse real-world environments, posing significant challenges.

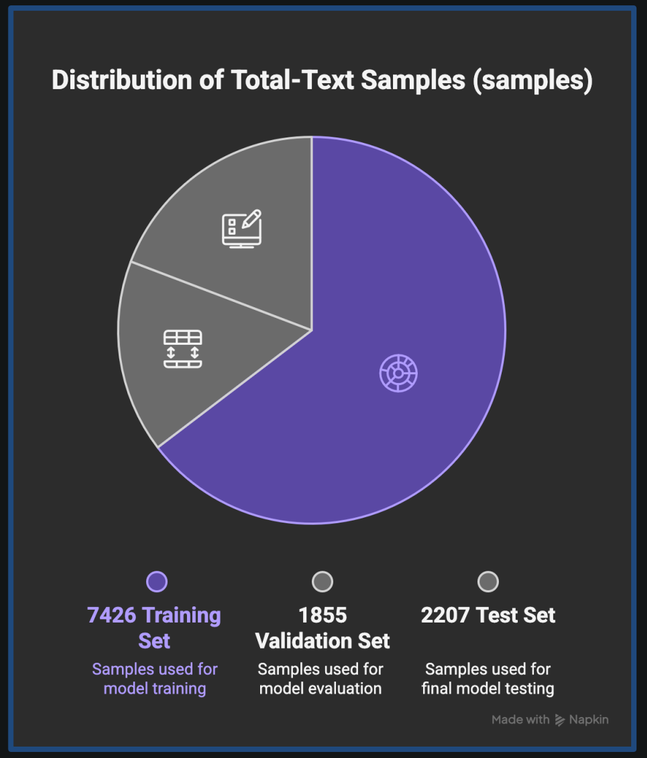
Scene Text Recognition on Images with Curved Text



Total-Text (Dataset)

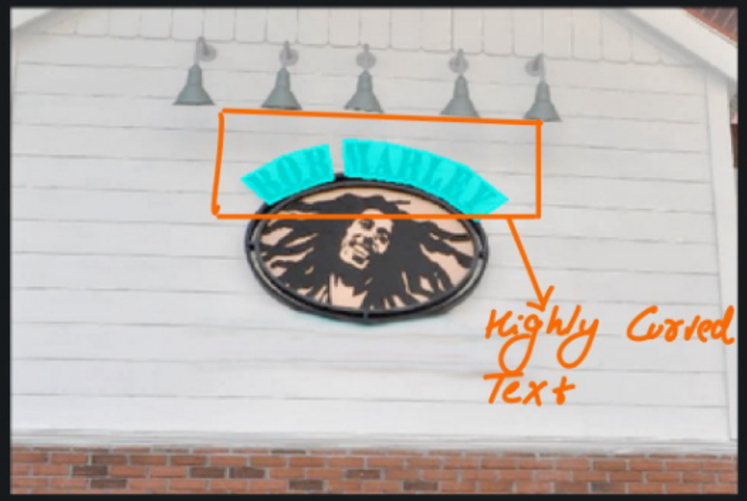
TotalText is one of the most widely used benchmark datasets for Scene Text Recognition (STR).
It contains highly diverse, real-world text images collected from natural scenes such as street signs, labels, boards, advertisements, objects, and handwritten surfaces.





Based on Crops

Samples


Highly Curved
Text


IDLE HAND
TATTOO
STUDIO

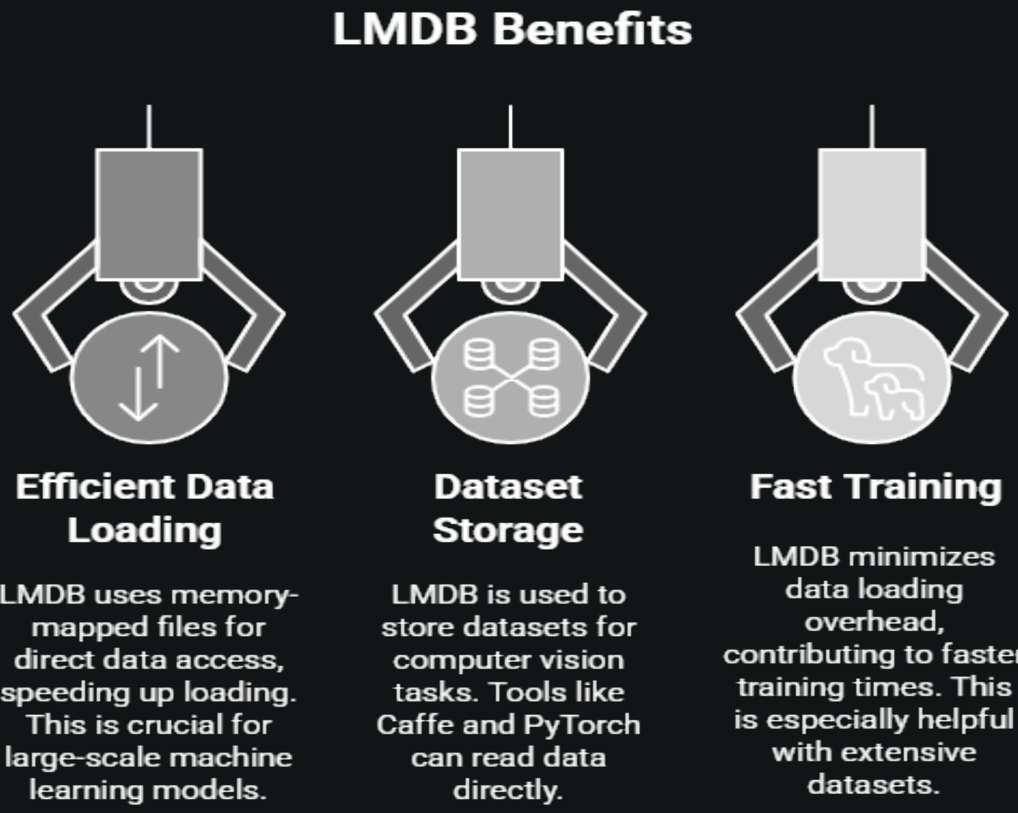

Multiple Text Crops
in a single Image
SERENBAN ICE CUBE
ICE
ICE
Other
Examples


Text with
low Brightness


BATTISTA'S
HOLE IN THE WALL

LMDB

LMDB (Lightning Memory-Mapped Database) is a high-performance, memory-mapped key–value store that delivers extremely fast, zero-copy reads with full ACID guarantees. It is widely used in AI pipelines and embedded systems due to its simplicity, low latency, and exceptional read scalability.



ABINet

CVPR(2021)

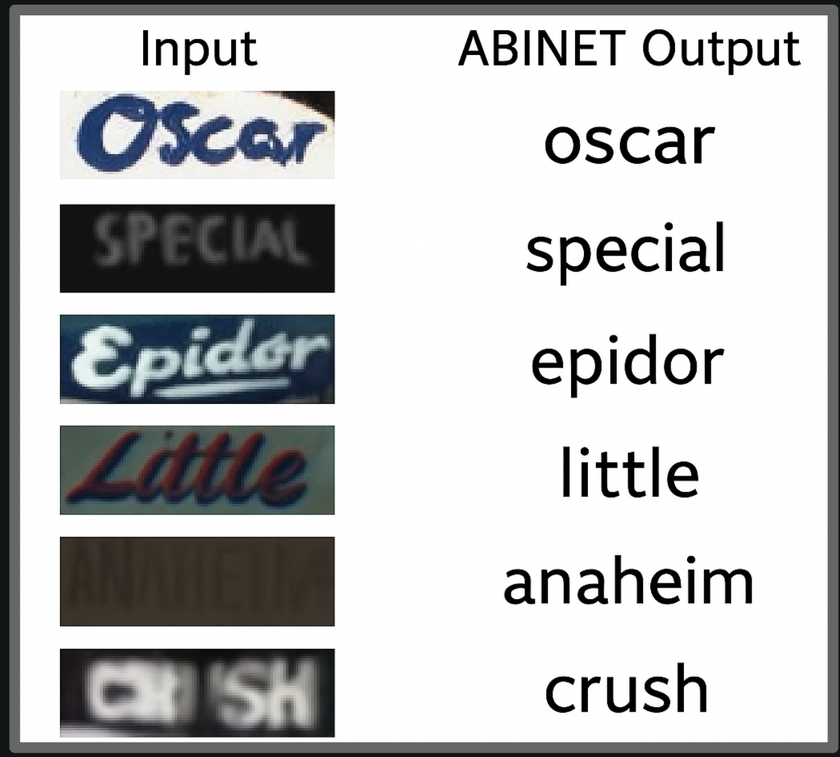
Vision-Language Integrated Model for Robust Scene Text Recognition

[Iterative Refinement, Bidirectional Transformer Language Model]

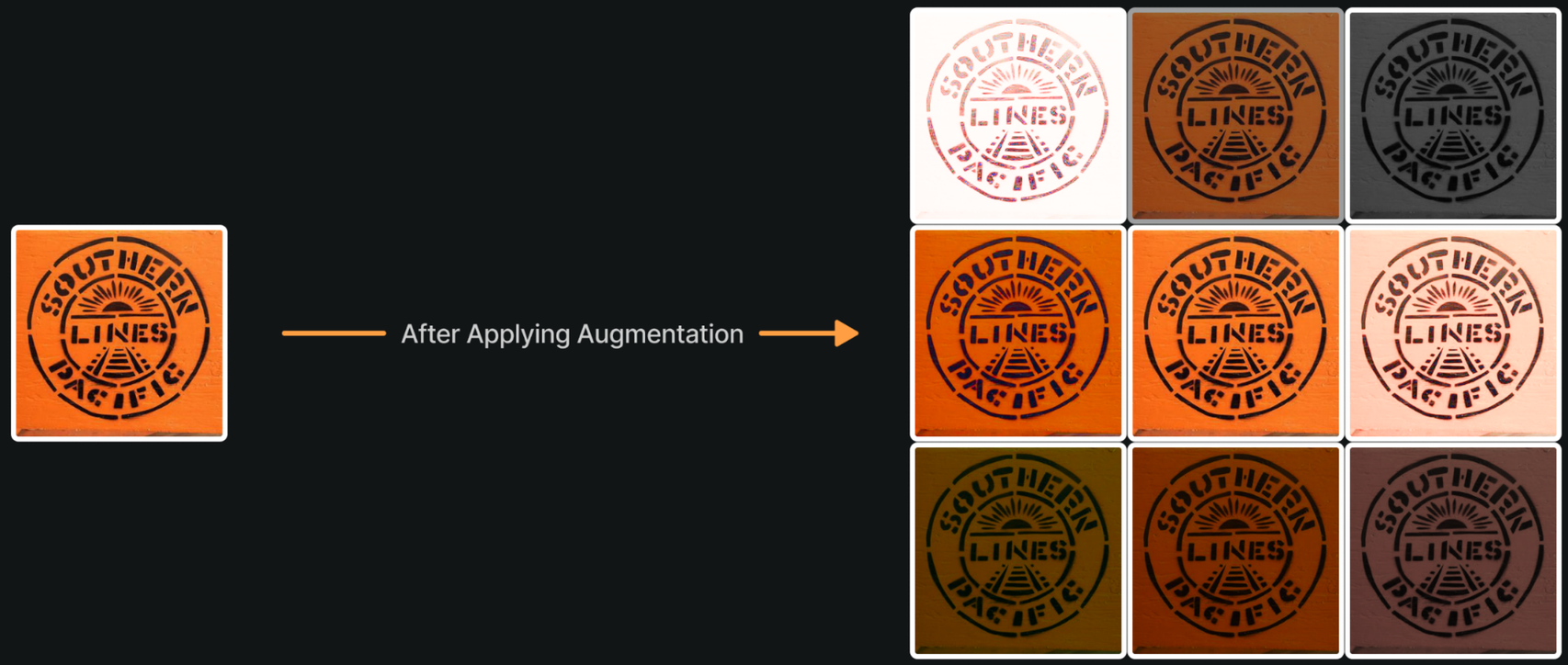
Why choose ABINet?

01 State-of-the-Art Performance [2021]

02 Vision–Language joint modeling



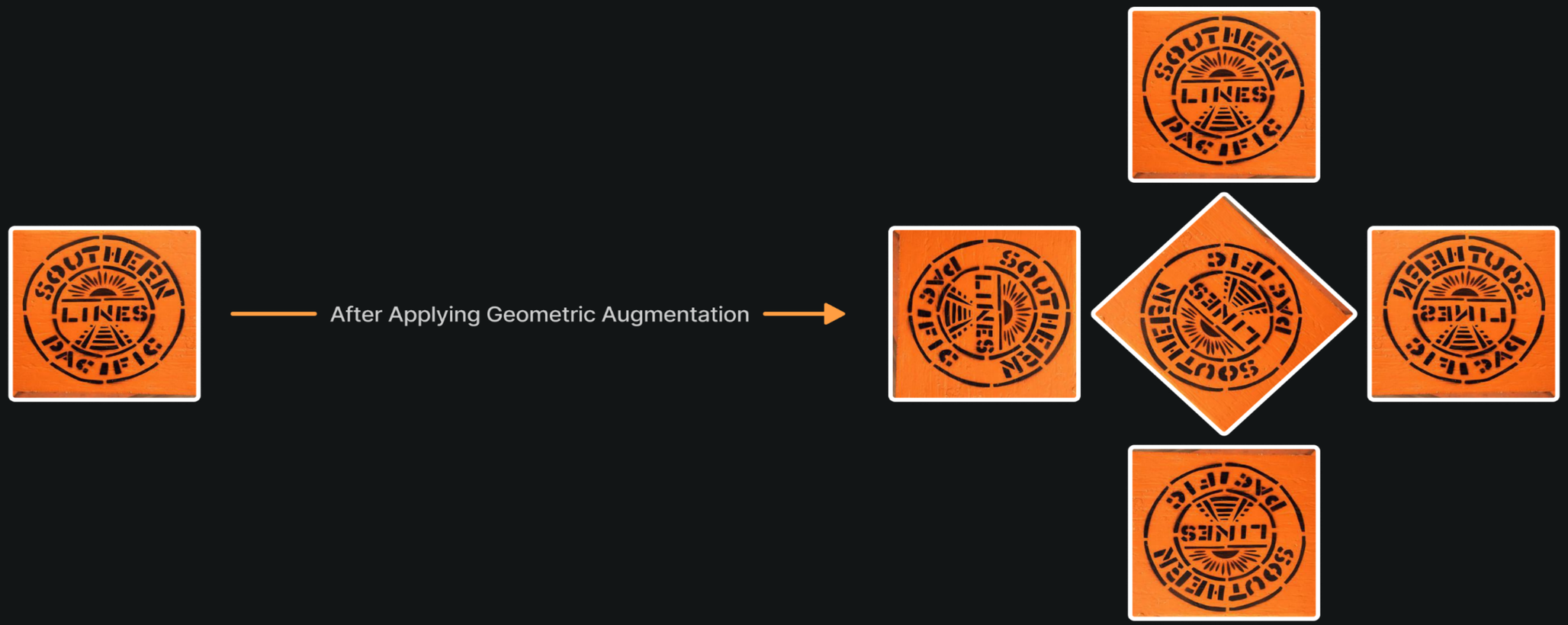
Augmentation



The Why?

TTA improves model accuracy by predicting on multiple augmented versions of the same test image and combining the results for a more robust final prediction.

Geometric Augmentation



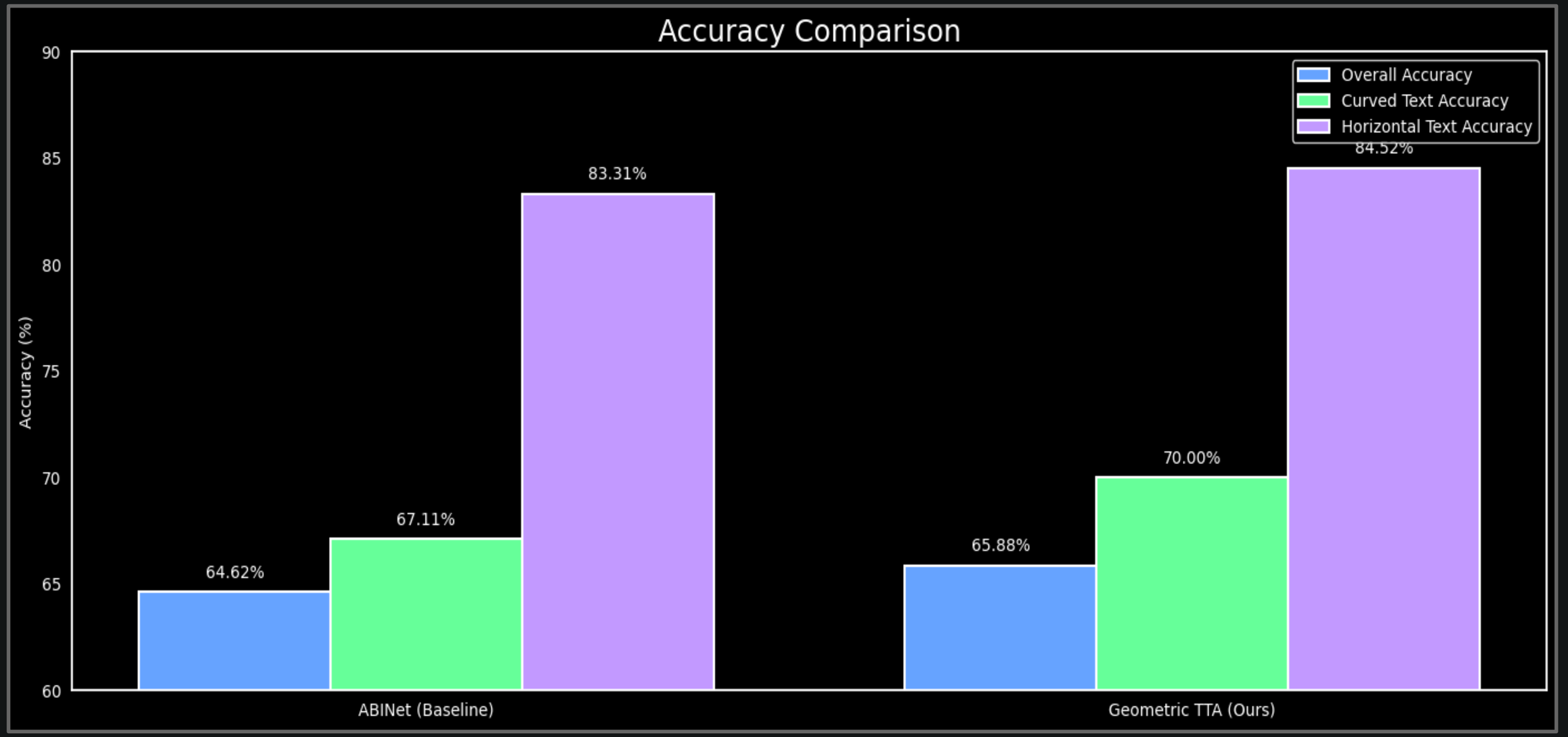
The Why?

Geometric TTA improves accuracy by generating spatially transformed versions of the test image—such as rotations or flips—and combining their predictions for more robust results.

Specifications of Test Time Augmentation

Feature	TTA (General Test Time Augmentation)	Geometric TTA
Type of Transformations	Includes *both* geometric + non-geometric (color, noise, blur, contrast, etc.)	Only geometric (rotation, scaling, perspective, crop, affine, flip)
Focus Area	Improves robustness to lighting, color, and noise variations	Improves robustness to shape, orientation, angle, and spatial distortions
Typical Augmentations	Color jitter, brightness, blur, noise, flips, crops	Rotation, affine transformations, perspective warp, scaling, flipping
Computational Cost	Higher (due to more types of augmentations)	Moderate (only spatial transforms)
Best Use Cases	General vision tasks (classification, detection)	OCR, Scene Text Recognition, Curved/ Rotated text

ABINet


Accuracy Comparison
Accuracy (%)
90
85
80
75
70
65
60
Overall Accuracy
Curved Text Accuracy
Horizontal Text Accuracy
64.62%
67.11%
83.31%
65.88%
70.00%
84.52%
ABINet (Baseline)
Geometric TTA (Ours)

Transition to PARSeq

Image Example	ABINet Baseline Prediction	ABINet + TTA (Ours)	PARSeq (Fine-Tuned)	Ground Truth
	PEA…	PEAK	PEAK	PEAK
	HEWBA	HEWBAC	CHEWBACCA	CHEWBACCA

PARSeq

ECCV (2022)

Scene Text Recognition with Permuted Autoregressive Sequence Models **by Darwin Bautista & Rowel Atienza.**

01 Inherent Geometric Robustness

02 Unified Contextual Decoding

PARSeq: Scene Text Recognition example. Input image with text 'CITY CAFE' (overelaid in red box) recognized by PARSeq as 'CITY CAFE'.

Source: https://arxiv.org/pdf/2208.09363

User	PARSeq
What text is in this image?	The text in the image reads 'CITY CAFE', as recognized by the PARSeq molel.
Is the model robust?	Yes, PARSeq has been shown to be robust to various distortions during text recogniti-on tasks.

PARSeq:

BaseLine

The PARSeq (Pre-Trained) model achieves a baseline accuracy of **93.02**% on the TotalText

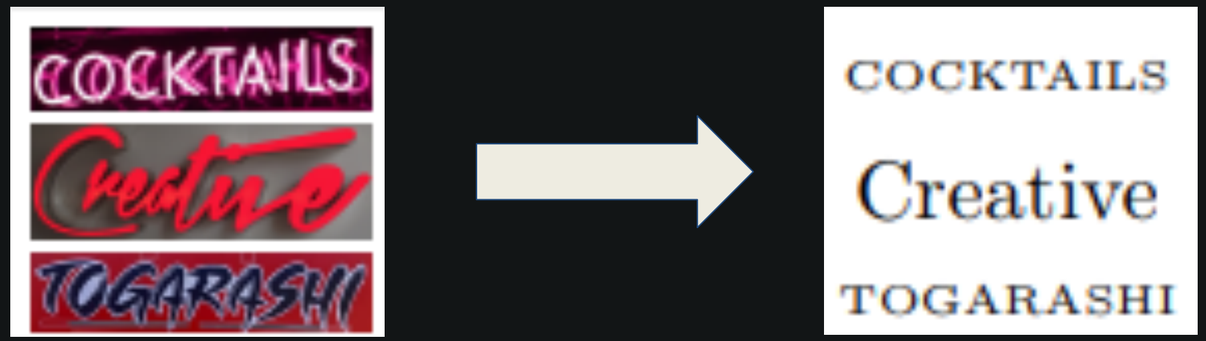
This serves as the reference performance against which all further improvements and adaptations were evaluated.

dataset.

Target

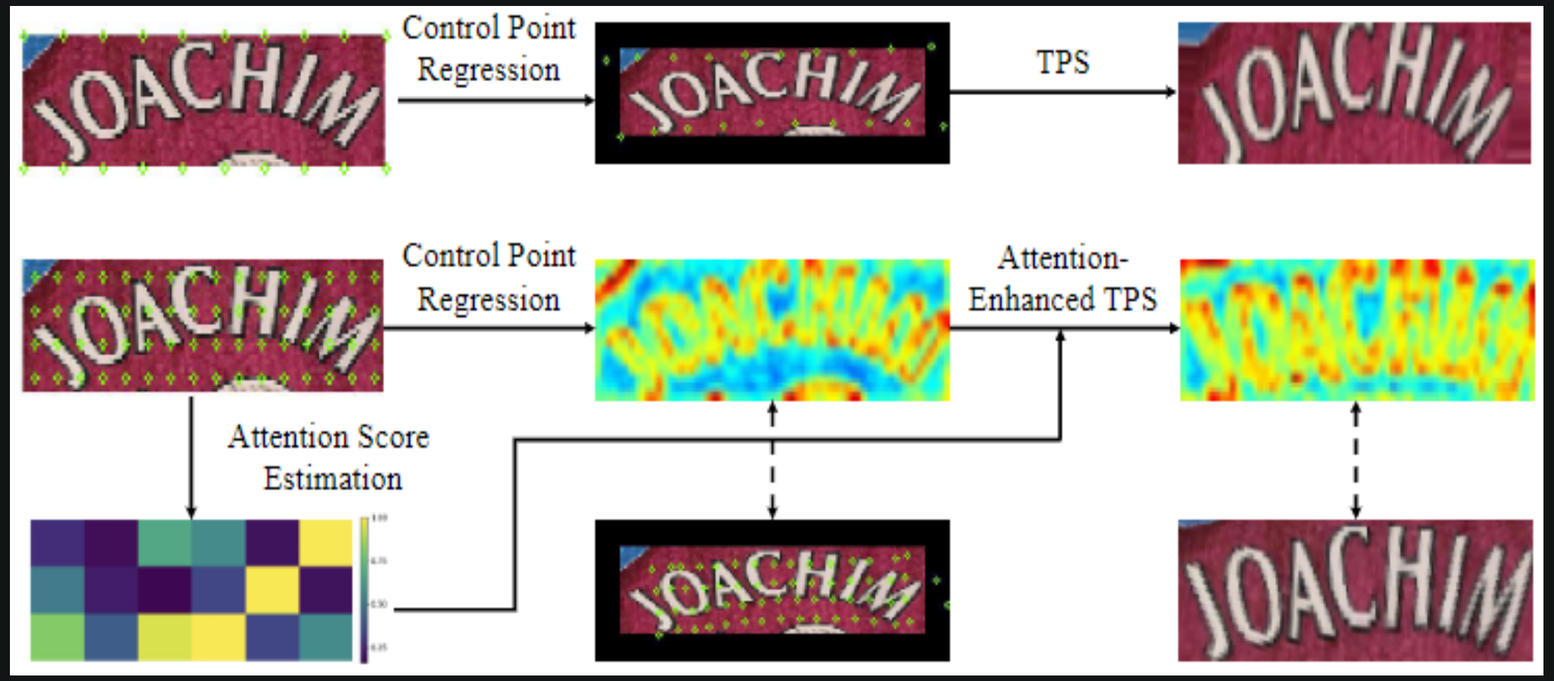
The primary goal of our project was to improve PARSeq’s accuracy on Total-Text scene-text images.

To accomplish this, we explored multiple architectural and training modifications, aiming to enhance the model’s ability to handle curved, irregular, and real-world text.



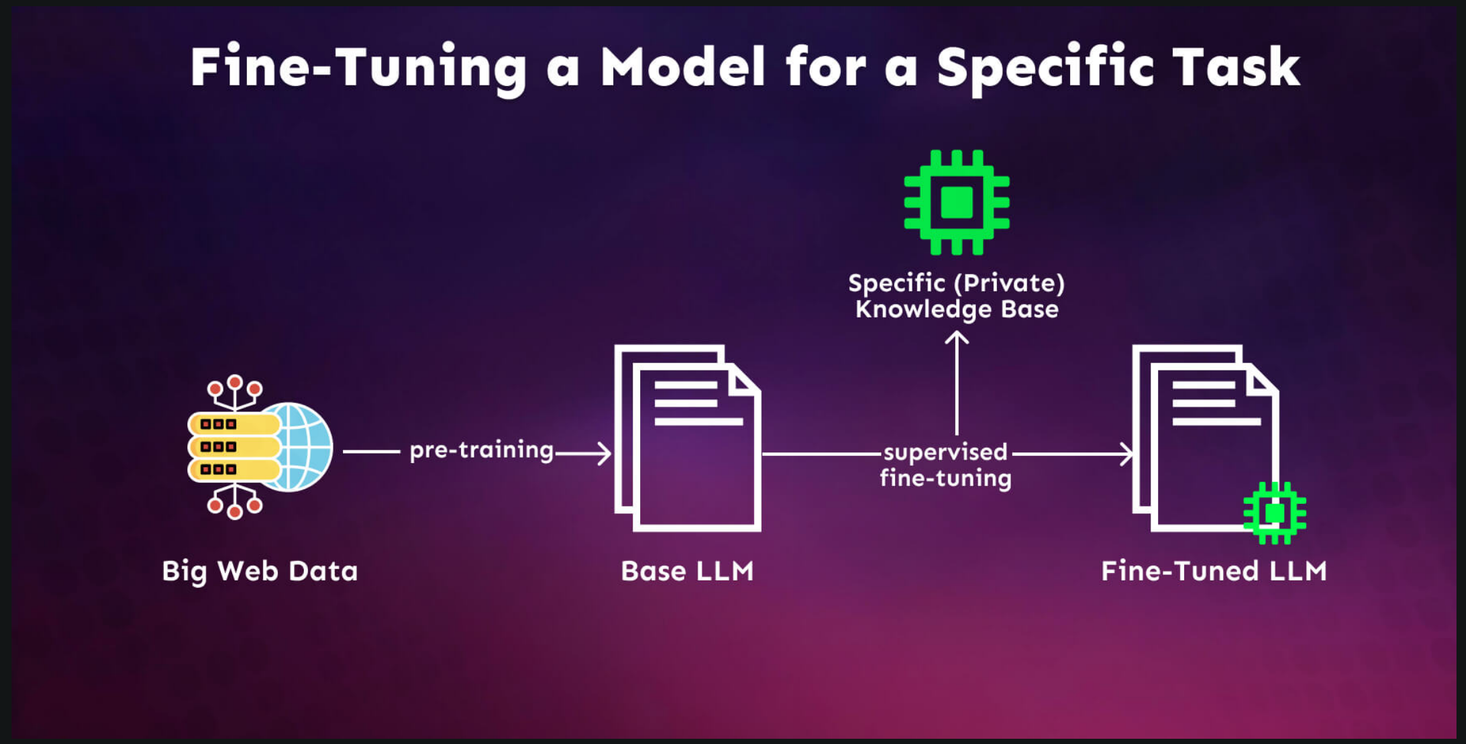
TPS (Thin Plate Spline):

TPS Rectification (Thin Plate Spline Rectification) is a geometric transformation technique used to correct distortions in an image—especially images containing text. It warps the image into a more uniform, rectangular, and readable form.



Fine Tuning :

Fine-tuning is the process of taking a pre-trained model and training it further on a task-specific dataset to improve performance for a particular domain or objective.



Adapter Modules (PEFT):

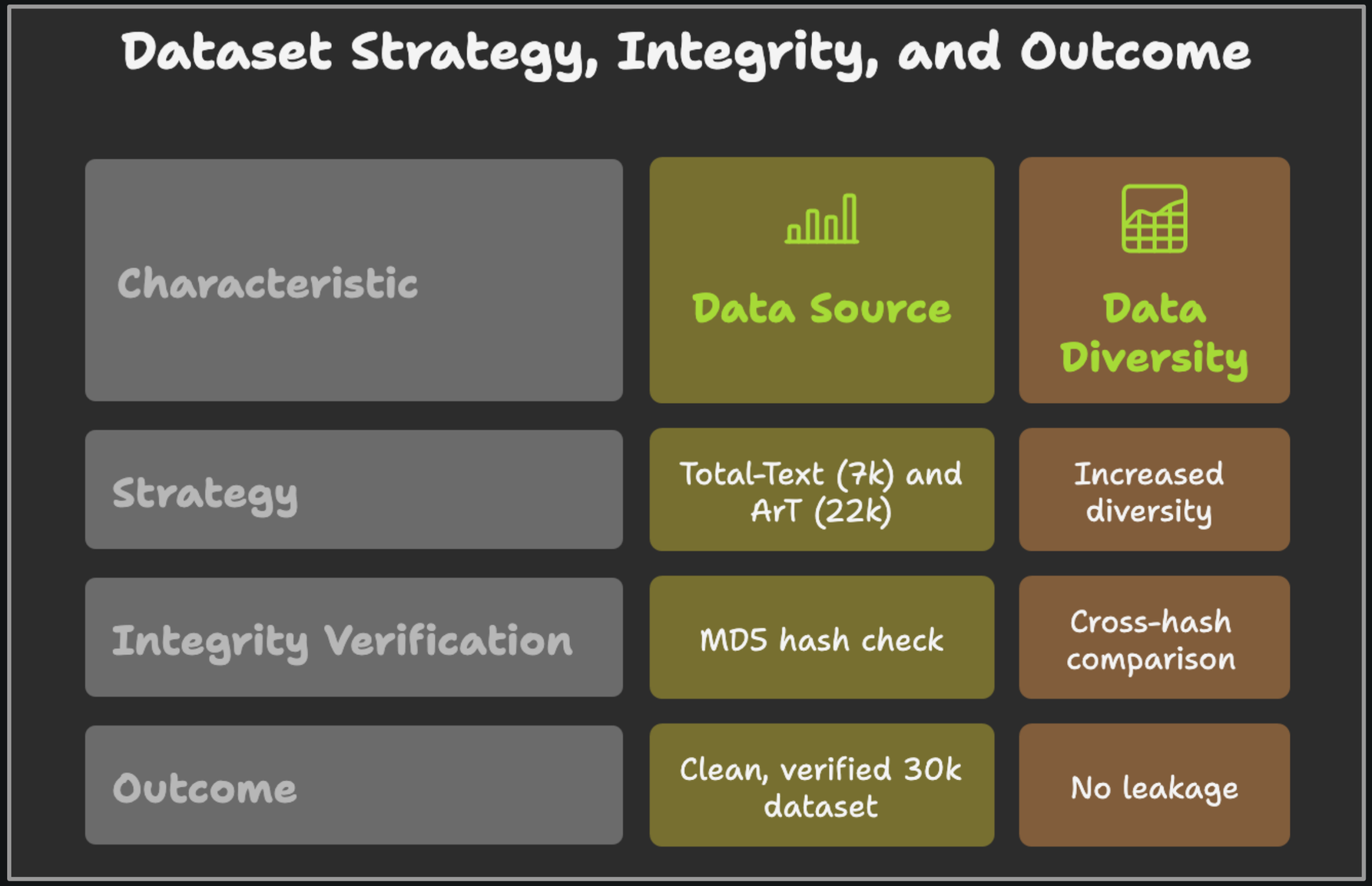
To enable parameter-efficient fine-tuning of the PARSeq Encoder by inserting lightweight, trainable bottleneck layers without modifying the backbone.

Architecture:

- Introduce a **down-projection** layer:
  $W_{down} \in \mathbb{R}^{d \times r}$
- Apply **non-linearity** (e.g., ReLU or GELU).
- Follow with an **up-projection**:
  $W_{up} \in \mathbb{R}^{r \times d}$
- Combine with the original hidden state using a **residual connection**:

$$h' = h + W_{up}\left(\sigma(W_{down}(h))\right)$$

Training Dataset

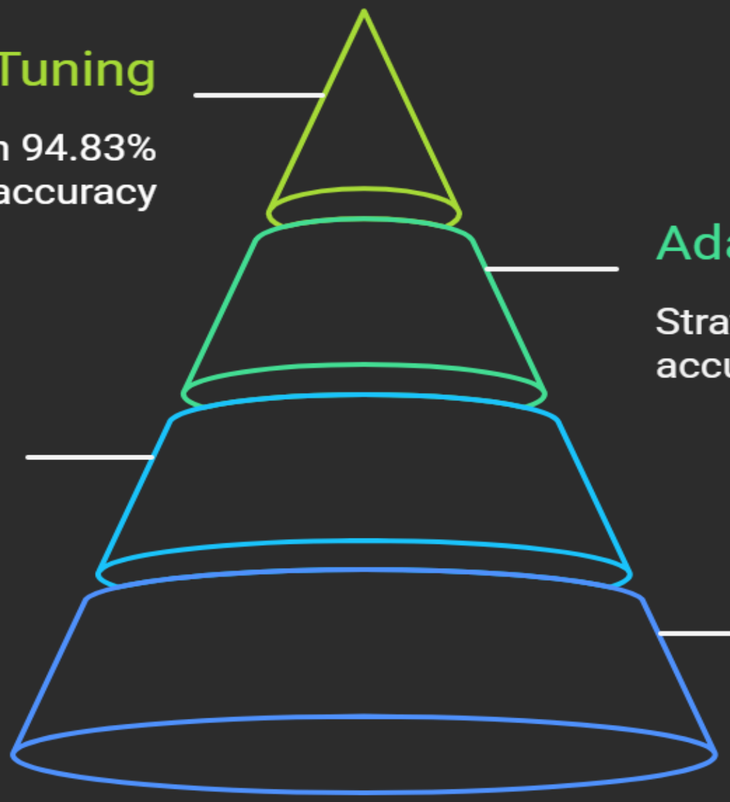


Quantitative Analysis

Comparison of Text Recognition Methods

Method	Total-Text Accuracy	Parameters	Notes
TRBA (Baek et al.)	87.1%	~50M	Standard ResNet+BiLSTM+Attn baseline
ABINet (Fang et al.)	89.4%	~37M	Language model based, strong on curved text
MATRN (Na et al.)	90.5%	~45M	Multi-modal attention
PARSeq (Official Baseline)	93.02%	23.8M	SOTA Permutation Autoregressive Sequence
Ours (Decoder Fine-tune)	94.83%	23.8M	-

Ablation



Decoder Fine-Tuning

Strategy with 94.83% accuracy

Adapter

Strategy with 94.20% accuracy

TPS Rectification

Strategy with 93.29% accuracy

PARSeq Baseline

Pre-trained model with 93.02% accuracy

Dataset-wise Performance

Model Comparison

Model	Baseline	Adaptor (fine-tuned base)	Fine-tuned Decoder	TPS (fine-tuned base)
Total-Text	93.02%	94.20% (+1.18%)	**94.83% (+1.81%)**	93.29% (+0.27%)
ArT	92.69%	92.13% (-0.56%)	**92.32% (-0.37%)**	90.74% (-1.95%)
LSVT	84.87%	82.75% (-2.12%)	**82.95% (-1.92%)**	78.99% (-5.88%)

Future Aspects

01 New Dataset Generation

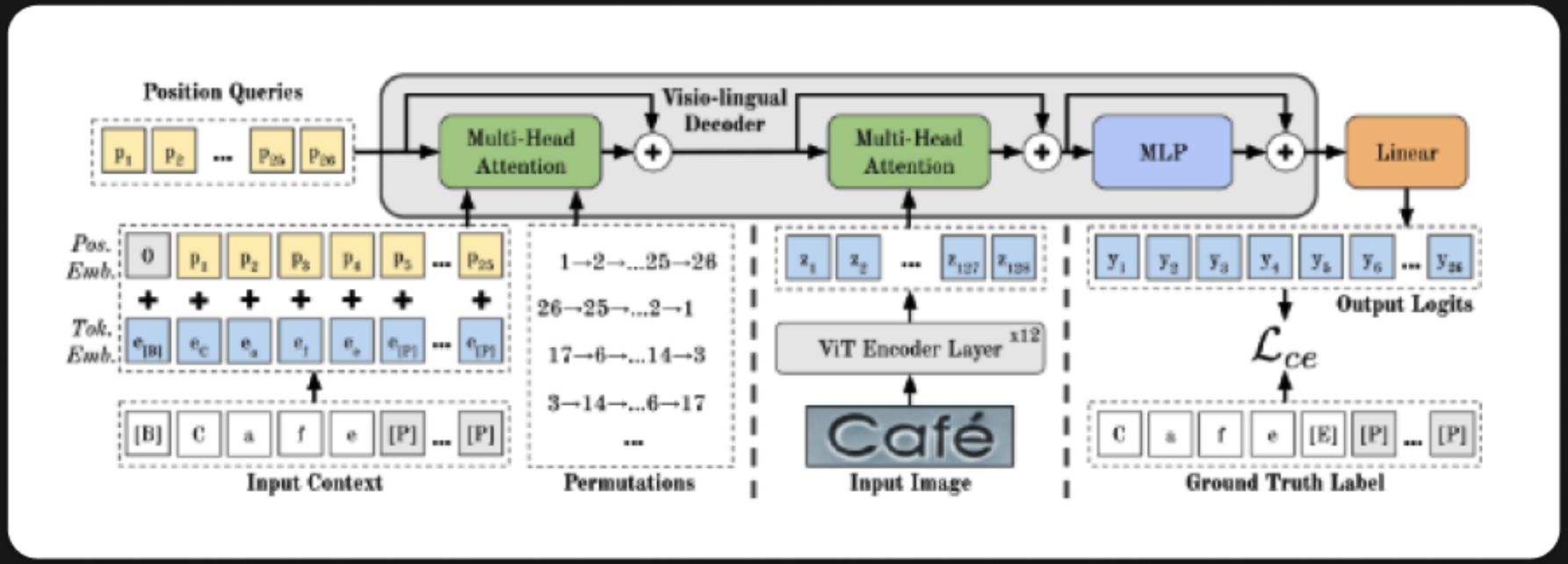
02 Vision-Specific LoRA

03 Cross-Modal Context

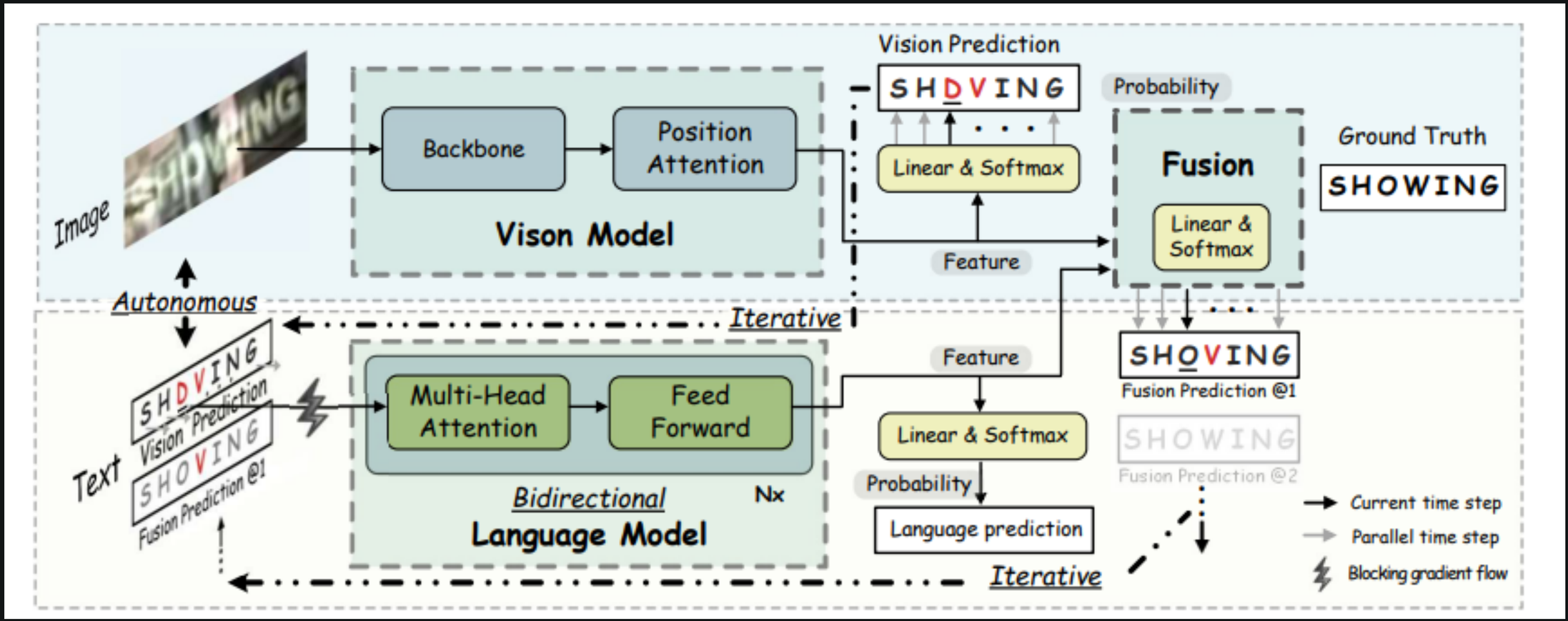
Thank You!

Demo Video

Appendix



PARSeq's Architecture



ABINet Architecture