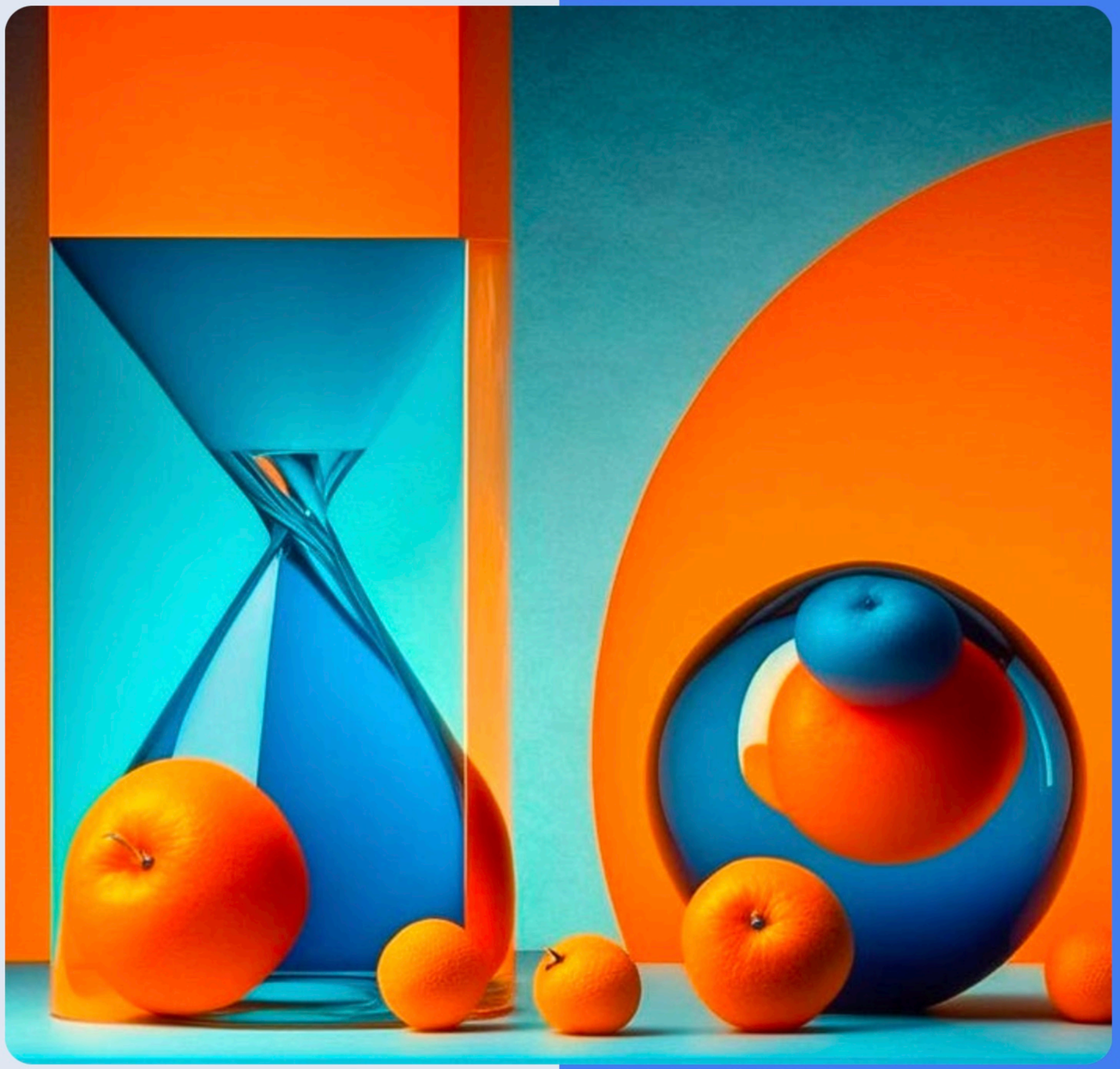
Image Caption Generator Using CNN & Vision Transformers



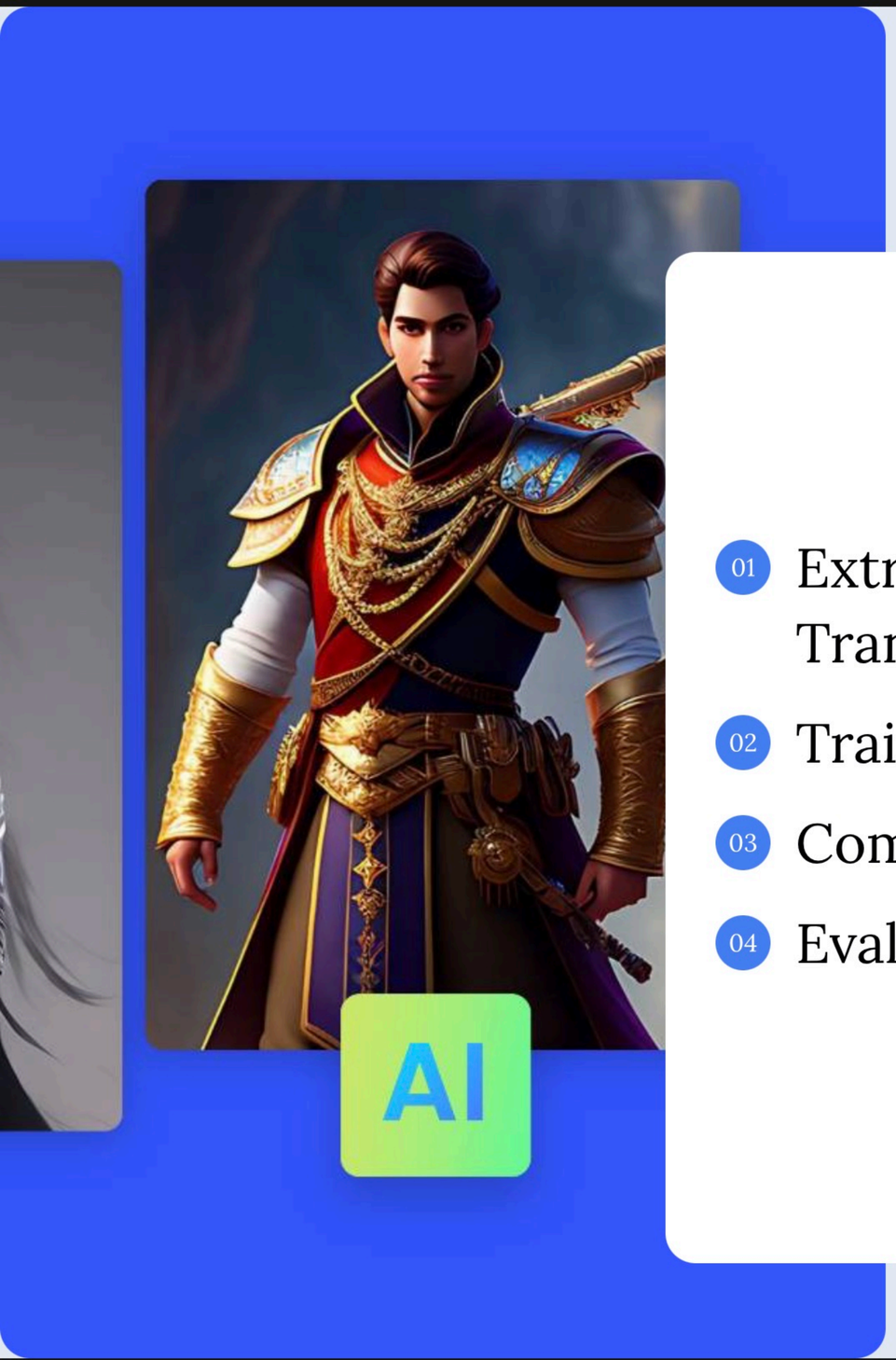
Project Overview

Combines **Computer Vision** & **NLP** to generate image captions

Explores **CNN-LSTM** vs **Vision Transformers**

Datasets: **MS COCO** & **Flickr_8K**

Goal: Generate **accurate**, **relevant** image descriptions



Objectives

01 Extract image features using CNNs and Vision Transformers

02 Train sequence models for caption generation

03 Compare performance across architectures

04 Evaluate using NLP metrics

Datasets

- MS COCO: Real-world images with rich annotations
- Flickr_8K: 8,000 images with five captions each
- Used for training and validating models

Methodology

Preprocess images & captions

Evaluate with metrics

Apply attention mechanisms

Extract features (CNNs & Vision Transformers)

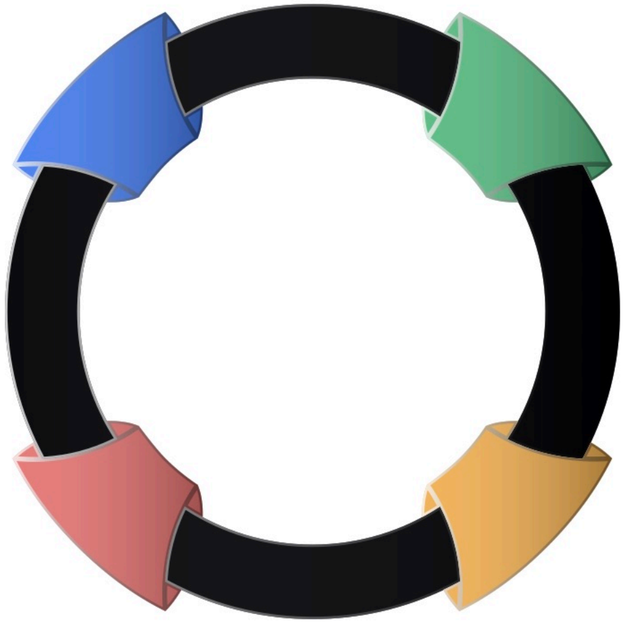
Tokenize captions and embed

Train LSTM-based sequence model

Feature Extraction

CNNs Used: ResNet-50, VGG-16

CNNs focus on local spatial features

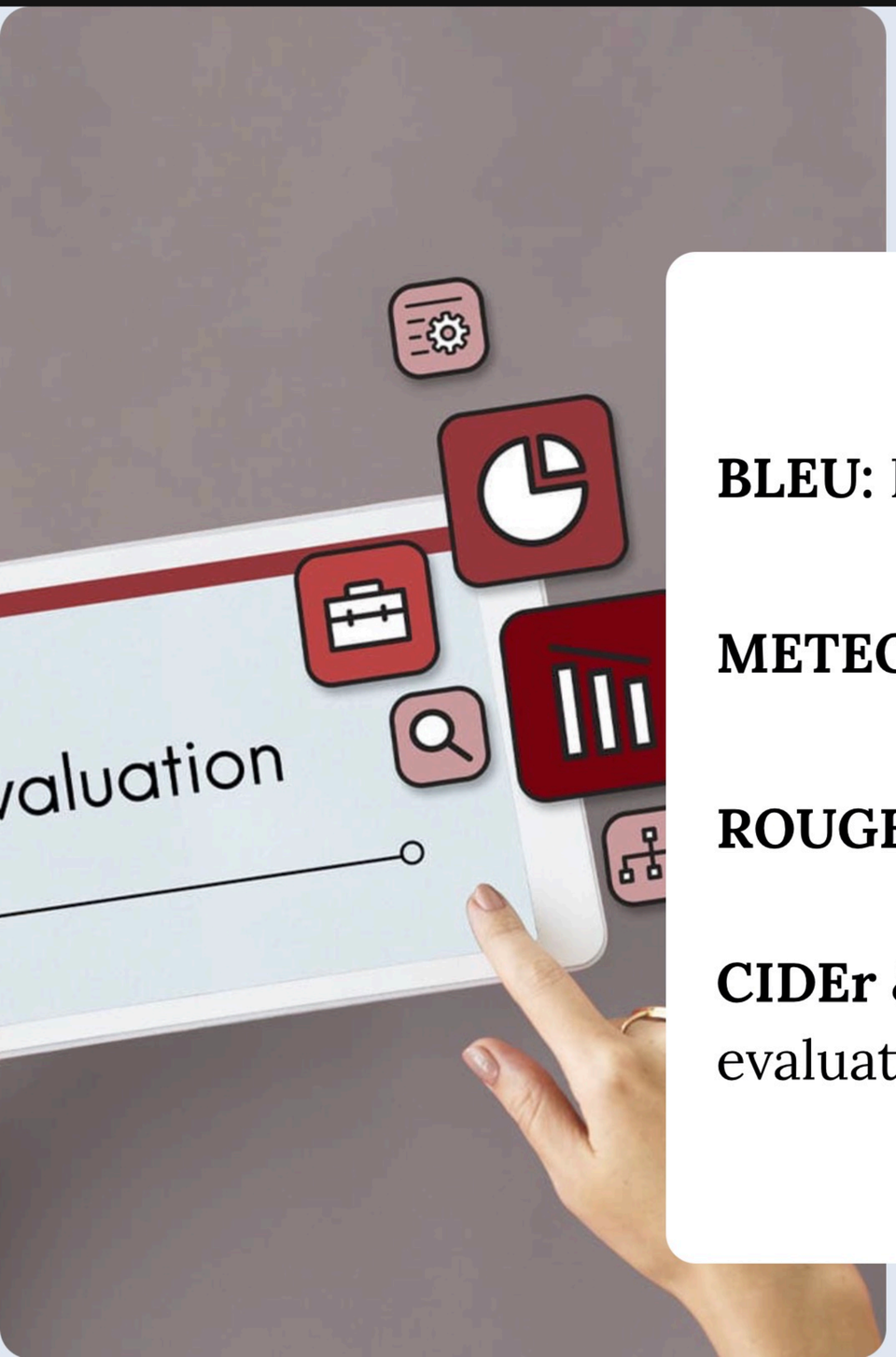


Vision Transformers: DINO, PVT, XCIT, SWIN

ViTs capture global features via self-attention

Caption Generation

- Captions tokenized and embedded
- LSTM used to predict sequences
- Attention used for better context and long-range dependency handling
- Training enhanced by fine-tuning & transfer learning



Evaluation Metrics

BLEU: N-gram precision

METEOR: Precision, recall, synonym matching

ROUGE: Overlap of longer sequences

CIDEr & SPICE: Semantic and consensus-based evaluation

Key Findings

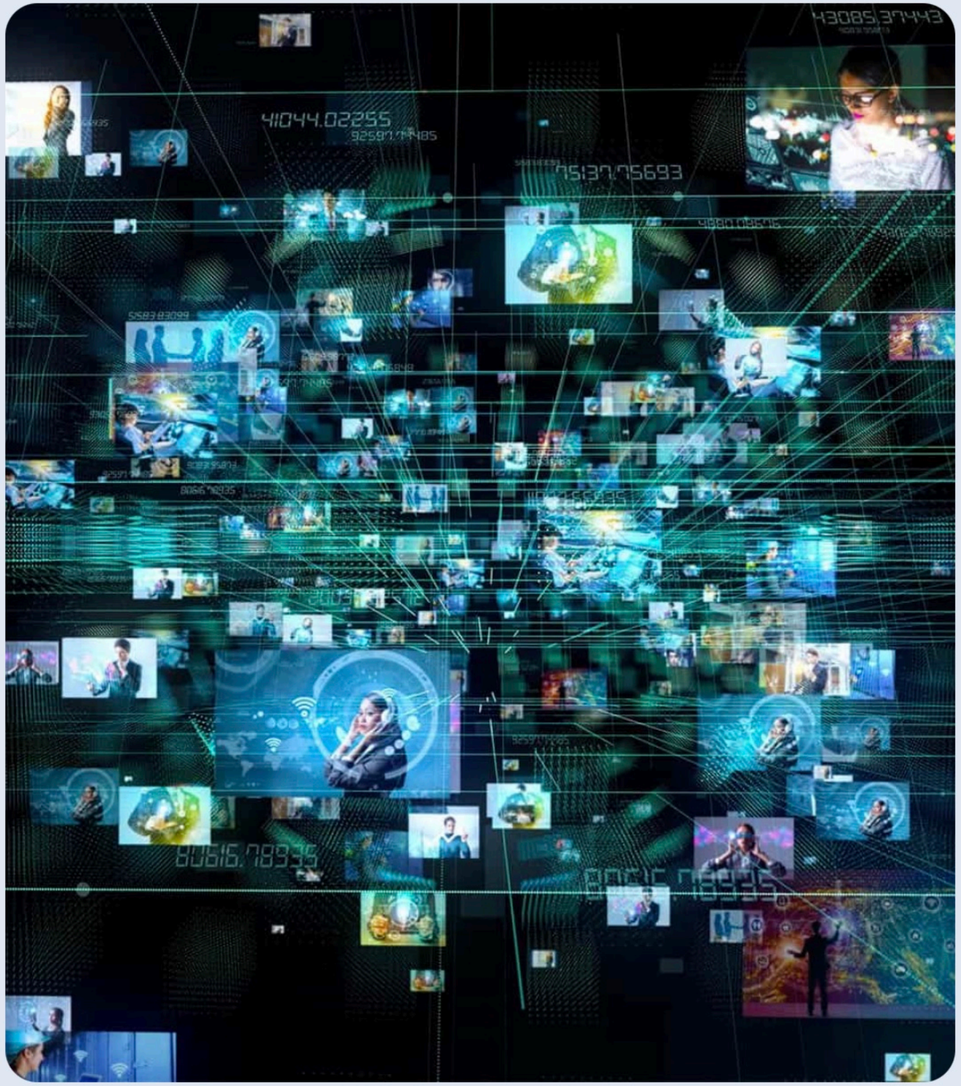
Vision Transformers outperform CNN-LSTM

SWIN Transformer provides best results

ViTs handle global dependencies more effectively

Improved accuracy and caption relevance

Conclusion



- Vision Transformers are more effective for image captioning
- Attention mechanisms improve descriptive accuracy
- Future scope: larger datasets, multimodal models, real-time applications