Image

A man wearing dark knee-length shorts and a white T-shirt with black symbol is holding a plastic bag

Text

Self Attention

Feedforward

FFN

Visual Encoder

Self Attention

Feedforward

FFN

Text Encoder

(a) Contrastive Visual Enhance(CVE)

Self Attention

CLS

VDL

r_t d_t s_t

Cross Attention

Reliable Candidate Space

Argmax

Rerank

Self Attention

EOT

TDD

[Mask] [Mask] [Mask]

(b) Discriminative Concept Calibrator(DCC)

Visual Dominant token Localization (VDL)

Local Attention Deviations

S_i

Self Attention

Text Dominant token Detect (TDD)

Self Attention

(symbol)

(holding)

(plastic)

&

Visual &Text Embedding

&

CLS Token & EOT Token

&

Visual Dominant Token & Text Dominant Token

s_t d_t r_t

Suppress & Decay & Retain