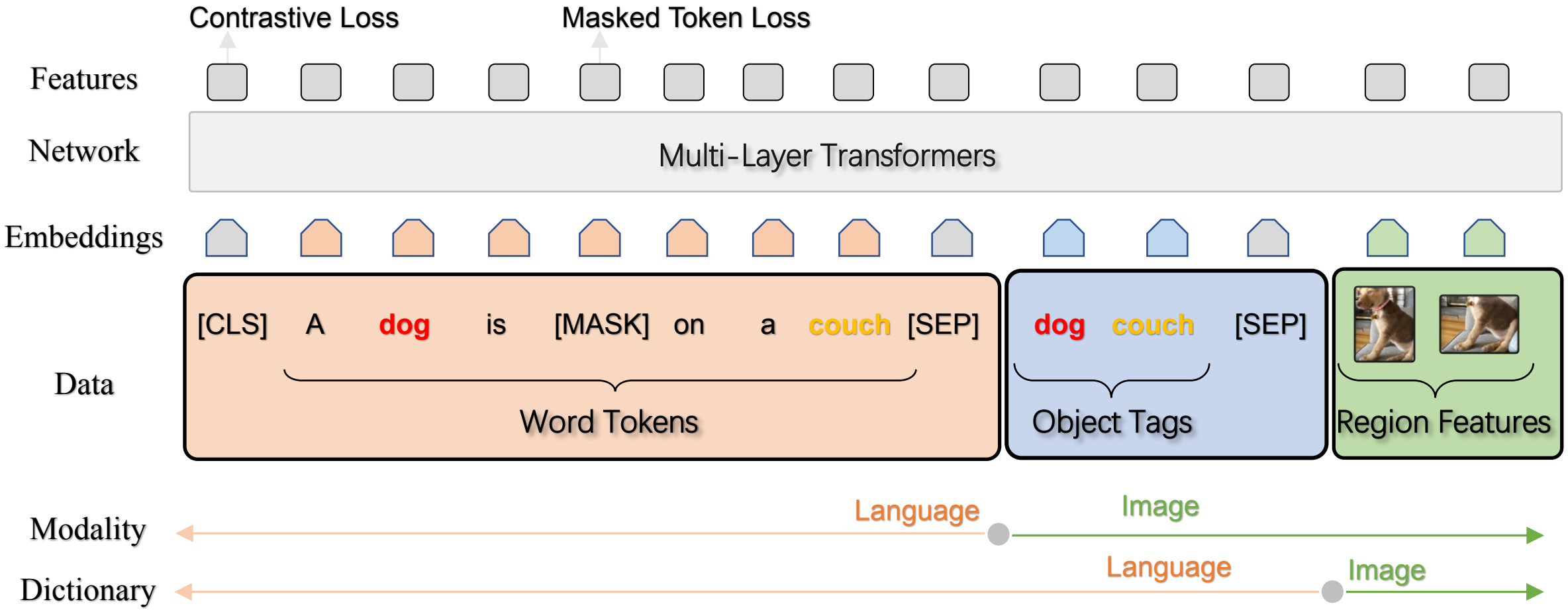
Contrastive Loss
Masked Token Loss
Features
Network
Multi-Layer Transformers
Embeddings
Data
[CLS] A dog is [MASK] on a couch [SEP]
Word Tokens
dog couch [SEP]
Object Tags
Region Features
Modality
Language
Image
Dictionary
Language
Image