Machine Learning Models Comparison

Model	Components and Purposes	Primary Loss Function	Attention Mechanism	Pre-training Objective	Input Transformation Method
BERT (text)	● Transformer Encoder: Contextual encoding	Cross-entropy	Bidirectional self-attention	Masked language modeling and Next sentence prediction	Learned token embeddings + positional encodings
GPT (text)	● Transformer Decoder: Autoregressive generation	Cross-entropy	Causal self-attention	Next token prediction	Learned token embeddings + positional encodings
BART (text)	● Transformer Encoder: Encoding corrupted input ● Transformer Decoder: Autoregressive reconstruction	Cross-entropy	Bidirectional (encoder) + Causal (decoder) self-attention	Denoising autoencoding tasks (e.g., text infilling, sentence permutation)	Learned token embeddings + positional encodings
T5 (text)	● Transformer Encoder: Task-specific encoding ● Transformer Decoder: Task-specific generation	Cross-entropy	Encoder-decoder attention + Self-attention	Span corruption	Learned token embeddings + relative positional encodings
CLIP (text/vision)	● Vision Transformer/ResNet: Image encoding ● Text Transformer: Text encoding	Contrastive loss	Self-attention (in Transformers)	Image-text alignment	Image: Patch embeddings or CNN Text: Token embeddings
DALL-E (text/vision)	● Discrete VAE: Image tokenization ● Transformer: Text-to-image generation	Cross-entropy	Causal self-attention	Autoregressive image generation conditioned on text	Text: Token embeddings Image: dVAE to discrete tokens

Model	Components and Purposes	Primary Loss Function	Attention Mechanism	Pre-training Objective	Input Transformation Method
Diffusion Models (vision)	● U-Net (often): Feature extraction and denoising ● Noise predictor: Estimates noise to be removed ● Conditioner: Guides generation process (optional)	Mean Squared Error (typically)	Self-attention (in U-Net) Cross-attention (for conditioning, if used)	Denoising score matching	Direct operation on pixel space + noise level embedding + optional conditioning signal
GAN (vision)	● Generator: Data generation ● Discriminator: Real/fake classification	Adversarial (minimax)	None (in basic GANs)	Distribution matching	Generator: Dense/Convolutional layers from noise vector Discriminator: Convolutional or dense layers
VQ-VAE (vision/audio)	● Encoder: Input encoding ● Vector Quantizer: Discrete representation ● Decoder: Reconstruction	Reconstruction loss + Vector Quantization loss	None (typically)	Discrete representation learning	Convolutional encoder to latent space, then vector quantization
SimCLR (vision)	● Data augmentation: Create positive pairs ● Base encoder (e.g., ResNet): Feature extraction ● Projection head: Representation for contrastive loss	Contrastive loss (e.g., NT-Xent)	None	Self-supervised visual representation learning	Convolutional layers (ResNet) + MLP projection head

Model	**Components and Purposes**	**Primary Loss Function**	**Attention Mechanism**	**Pre-training Objective**	**Input Transformation Method**
Wav2Vec (audio)	• Feature encoder: Raw audio embedding • Context network: Contextual representation • Quantizer: Discrete unit learning (in Wav2Vec 2.0)	Contrastive loss	Self-attention (in Wav2Vec 2.0)	Self-supervised audio representation learning	Convolutional layers for raw audio embedding
Whisper (audio)	• Encoder: Audio feature extraction • Decoder: Text generation	Cross-entropy	Encoder-decoder attention + Self-attention	Speech recognition and translation	Convolutional layers for audio, learned token embeddings for text