Model Comparison Report: CLIP vs. BLIP

This document gives a clear side-by-side comparison of two popular models used in multimodal AI: **CLIP** and **BLIP**. Both are designed to work with images and text but take different approaches under the hood.

Model	Text	Images	Audio	Mixed (Text + Image)
CLIP	✔	✔	✖	✔
BLIP	✔	✔	✖	✔

Architecture Overview

CLIP (Contrastive Language-Image Pre-training)

- Uses two separate encoders: one for images (ViT or ResNet) and one for text (Transformer).
- Both inputs are projected into the same embedding space using contrastive learning.
- Designed for comparing image-text pairs.

BLIP (Bootstrapped Language-Image Pre-training)

- Based on a more flexible encoder-decoder setup.
- Both image and text inputs go through their respective encoders, and then interact via cross-attention layers.
- Trained on multiple tasks: matching images and text, generating text from images, and contrastive learning.

Input Support

Both models work with text and images. Neither has native support for audio.

How They Work (Simple Diagram)

```
flowchart LR
    subgraph CLIP
        A1[Image] --> B1[Image Encoder (ViT/ResNet)]
        A2[Text] --> B2[Text Encoder (Transformer)]
        B1 --> C1[Shared Embedding]
        B2 --> C1
        C1 --> D1[Compare Similarity]
    end

    subgraph BLIP
        A3[Image] --> B3[Image Encoder (ViT)]
        A4[Text] --> B4[Text Encoder (BERT-style)]
        B3 --> C2[Cross-Attention]
        B4 --> C2
        C2 --> D2[Final Output: Retrieval, Captioning, QA]
    end
```

Typical Use Cases

Model	Best Suited For
CLIP	Zero-shot classification, image search, content filtering
BLIP	Image captioning, visual Q&A, image-text retrieval, dialog reasoning

How They Handle Multimodal Tasks

- **CLIP:**
  - Transforms images and text into a shared space.
  - Great for matching tasks and zero-shot applications.
- **BLIP:**
  - Uses attention to blend image and text information.
  - Can generate rich text outputs like captions and answers.

Quick Comparison

Feature	CLIP	BLIP
Architecture	Dual encoder (ViT + Transformer)	Encoder-decoder (ViT + BERT-like)
Input Types	Text, Images	Text, Images
Mixed Input Handling	Shared embedding space	Cross-attention-based fusion
Use Case Focus	Classification, search	Captioning, Q&A, reasoning
Generation Ability	Limited	Strong (text generation supported)
Training Data	400M web pairs	COCO, VG and bootstrapped captions
Audio Support	No	No
Strength	Fast, general-purpose, zero-shot	Versatile and generation-friendly

References

1. CLIP: viso.ai
2. CLIP vs BLIP: generativeai.pub
3. BLIP Paper (arXiv)
4. Salesforce Blog on BLIP