

# CLIP & Vision-Language — Quick Reference

The one idea, the loss functions, the model sizes, and the video recipe — on one page.

## THE ONE IDEA

A vision-language model puts images and text in one shared space of numbers (embeddings). Matching image-text pairs land close; mismatched pairs land far. Closeness = dot product of the two rows.

## HOW CLIP IS TRAINED

- Two encoders: an image encoder (Vision Transformer) and a text encoder (Transformer).
- Trained on 400 million image-caption pairs scraped from the web (OpenAI WIT dataset).
- Contrastive: for a batch of N pairs, score all N x N combinations; pull the diagonal up, push the rest down.
- Symmetric cross-entropy loss with a learned temperature. No human labels needed.

## ZERO-SHOT CLASSIFICATION

Label an image with category NAMES only, no training examples. Embed 'a photo of a {class}' for each label, embed the image, pick the closest. Best CLIP hit ~76% top-1 on ImageNet with zero ImageNet labels.

## MODEL SIZES (OpenAI ViT)

| Model               | Res. | Vision params | Embedding |
|---------------------|------|---------------|-----------|
| <b>ViT-B/32</b>     | 224  | 88M           | 512       |
| <b>ViT-B/16</b>     | 224  | 86M           | 512       |
| <b>ViT-L/14</b>     | 224  | 304M          | 768       |
| <b>ViT-L/14@336</b> | 336  | 304M          | 768       |

## FROM FRAME TO VIDEO

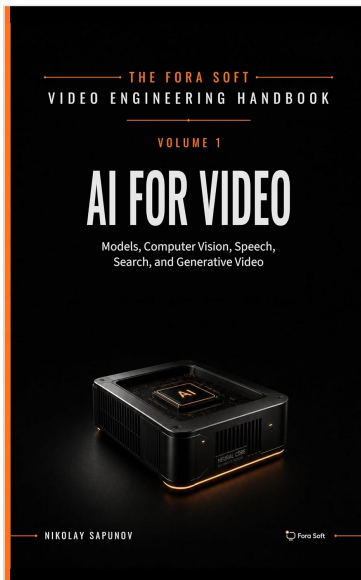
Sample frames ( $\approx 1/\text{sec}$ ), embed each with CLIP, average the embeddings into one clip embedding, match against text. 600 frames x 5 ms =  $\sim 3$  s GPU to make a 10-min clip searchable. X-CLIP / ViFi-CLIP add a temporal step when frame order matters.

## CLIP vs SigLIP

|                           | CLIP (2021)                | SigLIP / SigLIP 2                      |
|---------------------------|----------------------------|----------------------------------------|
| <b>Loss</b>               | Softmax (batch coupled)    | Sigmoid (each pair independent)        |
| <b>Big batches</b>        | Memory-heavy               | Cheap — fits larger batches            |
| <b>Zero-shot ImageNet</b> | $\sim 76\%$ (ViT-L/14@336) | $\sim 79\text{--}81\%$ (SigLIP 2 Base) |
| <b>2026 status</b>        | Legacy / compatibility     | Default in new open VLMs               |

## PITFALL

CLIP reads the gist, not fine detail. It is weak at counting, reading text, spatial layout, and negation ('no cars'). Use it for coarse matching, retrieval, and tagging — not precise reading.



GO DEEPER · VOLUME 6 OF 7

## AI for Video: A Complete Guide to Intelligence in Video Products

From Models and Inference to Computer Vision, Speech, Search, and Generative Video

By Nikolay Sapunov, CEO at Fora Soft · 1,444 pages · Kindle & paperback on Amazon

[Read it on Amazon →](#)
[amazon.com/dp/B0HH6TKPZ1](https://amazon.com/dp/B0HH6TKPZ1)

This reference covers how CLIP is trained, zero-shot classification, model sizes and the frame-to-video recipe. Volume 6 goes deeper into the capabilities built on it: search and RAG over a video archive, detection and tracking, generative video, and the deployment topologies that decide where embeddings run.

### WHAT YOU WILL FIND INSIDE

- Search and RAG over a video archive
- Detection, tracking and segmentation
- Generative video models
- Deployment topologies for embedding models

### THE FULL HANDBOOK · 7 VOLUMES · 9,000+ PAGES · ALL ON AMAZON

Vol. 1 · Video Encoding — pixels and color to H.264, HEVC, AV1, quality metrics and AI

Vol. 2 · Video Streaming I — network transport, contribution and ingest, HLS, MPEG-DASH, CMAF

Vol. 3 · Video Streaming II — ABR, CDNs, players, WebRTC, DRM and the economics of delivery

Vol. 4 · Audio for Video I — PCM, codecs, loudness, accessibility, real-time voice processing

Vol. 5 · Audio for Video II — jitter buffers, lip sync, Atmos, quality metrics, neural codecs

**Vol. 6 · AI for Video I — models and inference, computer vision, speech, search, generative video**

Vol. 7 · AI for Video II — real-time pipelines, agents, serving, cost, compliance, verticals

## Want natural-language search or auto-tagging across your video library?

We will embed your archive, measure retrieval quality on your own queries and ship search inside your product. 250+ video products shipped since 2005.

[Book a 30-min call →](#)
[WhatsApp →](#)
[Email us →](#)