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.

CLIP vs SigLIP

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

MODEL SIZES (OpenAI ViT)

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

FROM FRAME TO VIDEO

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

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.