Video Classification - Model layer

Model Layer — Video Classification

1) Transformer family (spatiotemporal attention)

VideoMAE (Masked Autoencoding for Video)

- **Why it matters:** Strong pretrained features via masked video reconstruction; efficient "tubelet" tokens; SOTA-ish fine-tuning on Kinetics/SSv2 with modest compute.
- **Checkpoints (HF):** `MCG-NJU/videomae-base-finetuned-kinetics`, `MCG-NJU/videomae-large-finetuned-kinetics`, `MCG-NJU/videomae-base`
- **Key ideas:** Tubelet embedding (3D patches), masked pretraining, temporal positional encodings, lightweight heads for classification.

TimeSformer (Divided Space–Time attention)

- **Why it matters:** Pioneering pure-ViT for video; factorizes attention into spatial+temporal for scalability.
- **Checkpoints (HF):** `facebook/timesformer-base-finetuned-k400`, `facebook/timesformer-base-finetuned-k600`
- **Key ideas:** Divided attention (space then time), ImageNet-style patch embeddings, standard ViT blocks extended to time.

X-CLIP (Video–Text contrastive)

- **Why it matters:** Zero-shot and few-shot action recognition by aligning videos with text prompts ("a video of …"); excellent when labels are scarce.
- **Checkpoints (HF):** `microsoft/xclip-base-patch32`, `microsoft/xclip-base-patch16`
- **Key ideas:** CLIP-style contrastive pretraining extended to video (frame sampling + temporal pooling), prompt engineering for classes.

---

2) ConvNet & hybrid families (strong baselines, efficient)

I3D / R(2+1)D / C3D (3D CNNs)

- **Why it matters:** Classic, reliable baselines; great for teaching and controlled ablations.
- **Checkpoints:** Often via PyTorchVideo/torchvision (exportable to HF Datasets pipelines).
- **Key ideas:** 3D convolutions (or (2+1)D factorization) to model time and space jointly.

SlowFast / X3D (meta-efficient 3D CNNs)

- **Why it matters:** High accuracy–efficiency trade-offs; dual-pathway (Slow for semantics, Fast for motion); X3D scales width/height/frames smartly.
- **Checkpoints:** PyTorchVideo (`slowfast_r50`, `x3d_m`, etc.).
- **Key ideas:** Multi-rate pathways (SlowFast), principled compound scaling (X3D).

Video Swin / UniFormer / MViT (hybrids)

- **Why it matters:** Windowed or multiscale attention with Conv-like inductive bias; strong accuracy/latency.
- **Checkpoints:** Common in MMAction2/PyTorchVideo ecosystems; some ports exist on HF Hub.
- **Key ideas:** Hierarchical tokens, windowed attention (Swin), multiscale attention (MViT), Conv–Attention fusion (UniFormer).

---

3) Multimodal variants (optional but powerful)

Audio–Visual models (AVSlowFast, fused Transformers)

- **Why it matters:** Actions with salient sounds (e.g., musical instruments, speech) benefit from audio fusion.
- **How:** Concatenate or cross-attend visual tokens with log-mel spectrogram features.

Video–Text (zero-shot)

- **Why it matters:** Open-vocabulary classification; great when class taxonomy changes often (e.g., social media trends).
- **Examples: X-CLIP** (above), CLIP-based video pooling heads.

Architectural innovations (what to teach and why)

- **3D Conv vs (2+1)D factorization:** 3D Conv models motion directly; (2+1)D reduces parameters by separating spatial and temporal convs.
- **Divided space–time attention (TimeSformer):** Improves scalability by factorizing attention; helps on long clips.
- **Tubelet tokens (VideoMAE/ViViT):** 3D patches lower token count vs per-frame patches → faster training/inference.
- **Masked video pretraining (VideoMAE):** Learns robust motion/appearance features without labels → superior fine-tuning.
- **Dual-pathway (SlowFast):** Explicitly models fast-changing motion and slow semantics.
- **Multiscale attention (MViT/Video Swin):** Hierarchical tokens and pooled attention handle long contexts efficiently.
- **Contrastive video–text (X-CLIP):** Open-vocabulary actions via prompts; strong zero-shot transfer.
