### **1. SSCD (Self-Supervised Copy Detection – Meta AI)**

[https://github.com/facebookresearch/sscd-copy-detection](https://github.com/facebookresearch/sscd-copy-detection?tab=readme-ov-file)

* **Strengths:**
  + **Speed & Scalability:** Generates compact embeddings (512 dimensions) per image and can process hundreds of frames per second using a GPU.
  + **Robustness:** Trained with strong augmentations to handle modifications like cropping, color changes, overlays, and more.
  + **Ease of Use:** Pre-trained and ready to integrate into existing pipelines.
* **Weakness:**
  + **Granularity:** Being a global descriptor, it might miss very small or localized copied segments within a frame.
* **Ideal For:** Large-scale, fast indexing where overall robustness is key.

### **2. VSAL (Video Similarity and Alignment Learning)**

<https://github.com/PVCD-VSAL/vsal>

* **Strengths:**
  + **Precision:** Uses a frame-by-frame similarity analysis and alignment strategy to not only detect but also precisely localize the copied segments.
  + **Benchmark Performance:** Achieves state-of-the-art results in standard benchmarks like VCDB and FIVR-200k.
* **Weakness:**
  + **Computational Cost:** More resource-intensive as it requires comparing frames and running an alignment network, typically needing a GPU for reasonable performance.
* **Ideal For:** Scenarios where pinpointing the exact copied segments is critical, even if it means slower processing.

### **3. ViT + SAM (Tencent VSC 2023)**

<https://github.com/facebookresearch/segment-anything>

<https://github.com/Jarus77/SAM-ViT-Model>

* **Strengths:**
  + **Accuracy:** Delivers top-notch performance (with reported mAP around 87% and segment localization around 91%), even under multiple transformations.
  + **Advanced Pipeline:** Combines a Vision Transformer for robust embeddings with a specialized alignment model (SAM) to refine detection.
* **Weakness:**
  + **Complexity & Hardware Requirements:** The multi-stage pipeline is complex to implement and requires high-end GPUs (like A100) for optimal performance.
* **Ideal For:** Use cases where the highest detection accuracy is needed and the available infrastructure can support the computational demands.

### **4. Other Approaches and Multimodal Techniques**

* **VCDT:** A Transformer-based method that shows promise but is currently more experimental and less plug-and-play.
* **Multimodal Solutions:** Integrating audio fingerprinting (using models like VGGish) alongside video analysis can catch cases where only one modality is reused, reducing both false negatives and false positives.

**5. Hackathons**

### [https://github.com/drivendataorg/video-similarity-challenge](https://github.com/drivendataorg/video-similarity-challenge?tab=readme-ov-file)

### **Final Consideration: A Hybrid Approach**

No single model is universally “best” for every scenario. A practical strategy might involve:

1. **Candidate Filtering:** Start with a fast global descriptor like SSCD or a ViT-based model to quickly narrow down potential matches.
2. **Precise Verification:** Use a more precise method like VSAL or the ViT + SAM pipeline to confirm and localize the copied segments.
3. **Multimodal Integration:** Add an audio analysis component to handle cases where only the audio track is reused or modified.

**Conclusion:**

* For large-scale, quick scanning, **SSCD** is highly effective.
* For precise detection and localization, **VSAL** or **ViT + SAM** are preferable, with the latter offering state-of-the-art performance at the cost of higher complexity.
* A hybrid system combining these methods, possibly with an audio fingerprinting layer, would likely provide the most robust solution overall.