Lightweight Multi-Modal Fusion for Robust Video Retrieval: When Action Features Boost Confidence, Not Detection

Denis Billi
Independent Researcher
denis@denisbilli.it

November 2025

Abstract

Video content retrieval systems must handle realistic perturbations like cropping, compression, and speed changes while maintaining semantic accuracy. We present **StoryHash**, a lightweight multi-modal fusion architecture combining visual (CLIP), temporal (action features), and structural (scene graph) embeddings into a 527-dimensional vector. Through systematic evaluation on 89 DAVIS videos with 14 transformation types (1,242 tests), we achieve **98.6% Recall@1** and **0.9748 average similarity**. Our ablation study reveals a surprising dichotomy: **action features boost confidence (+28% similarity) but minimally impact detection rate (+0.4% Recall@1)**. This finding validates efficiency-focused designs where lightweight temporal features (15D, 2.8% overhead) dramatically improve match confidence without requiring expensive temporal models. We identify a CLIP vertical flip bias (84.3% detection) and show rotation is the only transformation where action features aid detection (-2.3% degradation without). Our work provides quantitative evidence for complementary fusion architectures where visual features provide discrimination and temporal features provide confidence.
Keywords: video retrieval, multi-modal fusion, robustness, ablation study, CLIP, action recognition

1 Introduction

1.1 Motivation

Video content platforms face a persistent challenge: detecting modified or re-uploaded content despite realistic perturbations (cropping, compression, speed changes, flipping). Applications span copyright protection, content moderation, and tamper detection. While deep learning models excel at semantic understanding, their robustness to real-world transformations remains understudied, particularly when combining multiple modalities.

1.2 Research Questions

Our investigation centers on three fundamental questions that bridge theoretical understanding and practical deployment of multi-modal video retrieval systems.

First, we examine robustness: Can a multi-modal video fingerprint achieve greater than 95% detection rate under realistic perturbations encountered in social media ecosystems? Existing benchmarks focus on clean academic datasets (UCF101, Kinetics), but real-world content undergoes

aggressive compression, cropping, speed adjustments, and geometric transformations during viral sharing. We hypothesize that fusing complementary modalities (visual semantics, temporal motion, structural relationships) will provide redundancy against perturbations that affect different feature types asymmetrically.

Second, we investigate modality contribution: How do visual, temporal, and structural features contribute to detection capability versus match confidence? Prior multi-modal fusion work treats all modalities as equally necessary for performance, optimizing joint accuracy on action recognition or video captioning tasks. We challenge this assumption by isolating each modality's impact through ablation: does removing temporal features cause catastrophic failure (indicating necessity) or graceful degradation (indicating complementarity)? Crucially, we distinguish between detection rate (Recall@k: whether the system finds matches) and confidence (similarity scores: how certain the system is about matches)—a distinction critical for automated decision-making in copyright and moderation contexts.

Third, we quantify efficiency trade-offs: What is the cost-benefit of adding lightweight temporal features to visual embeddings? Temporal models like VideoMAE and I3D extract rich spatiotemporal representations (768-2048 dimensions) but require expensive 3D convolutions and large memory footprints. We explore whether minimal motion statistics (15 dimensions aggregated from object tracking) can provide practical value—improved confidence, robustness to temporal perturbations—without matching the computational complexity of full temporal architectures. This question addresses deployment realities: production systems need "good enough" solutions that run at scale on CPU hardware, not state-of-the-art models requiring GPU clusters.

1.3 Contributions

This work makes five key contributions to video retrieval and multi-modal fusion research.

Comprehensive robustness benchmark: We conduct systematic evaluation of video fingerprint robustness through 1,242 tests spanning 14 transformation types that simulate real-world content modifications. Our benchmark includes spatial perturbations (crop 10%/20%, rotate 5°/10°), photometric changes (brightness +20%/+40%, compression 50%/25% bitrate), temporal manipulations (speed 0.9×/1.1×), geometric transformations (flip horizontal/vertical), content additions (watermark overlay), and combined multi-step perturbations (crop + flip + speed) mimicking messaging app forwarding. Unlike prior work focusing on additive noise or blur, our transformations reflect social media ecosystems (TikTok speed adjustments, WhatsApp compression, platform watermarking) where copyright evasion and viral sharing occur.

3-way ablation study: We quantify each modality's contribution through controlled ablation, testing Fusion (CLIP 512D + Action 15D + Graph 8D), CLIP-only (512D visual), and Action-only (15D temporal) configurations on identical test sets. This isolation reveals which modalities provide detection capability versus match confidence—a distinction absent in prior multi-modal fusion work that reports only aggregate accuracy metrics. Our methodology enables architectural decisions grounded in empirical cost-benefit analysis rather than intuition about modality complementarity.

Novel finding on action features as confidence boosters: Our ablation uncovers a surprising dichotomy: action features minimally improve detection rate (+0.4% Recall@1, from 98.2% to 98.6%) but dramatically boost match confidence (+28% average similarity, from 0.70 to 0.97). This finding challenges the assumption that all modalities must contribute equally to detection, instead validating efficiency-focused architectures where lightweight features (15D, 2.8% overhead) enhance certainty in matches already found by strong visual backbones. The confidence improvement has practical implications: copyright systems gain legal justification for automated takedowns, content moderation reduces human review load, and real-time retrieval achieves higher precision at scale.

CLIP vertical flip bias identification: We discover and quantify a systematic orientation bias in CLIP: vertical flip reduces detection to 84.3% while horizontal flip achieves 100%, a 16-percentage-point gap. Through failure case analysis, we trace this to CLIP's training data distribution—400M internet images are overwhelmingly "upright," causing the model to learn gravity-dependent features. This bias creates adversarial vulnerability (bad actors can exploit flip_v for 16% evasion without perceptual quality loss) and informs mitigation strategies (fine-tuning with orientation augmentation). Our finding contributes to the growing literature on dataset bias in foundation models, demonstrating how training distribution shapes robustness profiles.

Open-source implementation: We release the complete StoryHash pipeline and evaluation code, including feature extraction (CLIP, SAM2 tracking, scene graphs), FAISS indexing, robustness benchmark scripts, and ablation analysis tools[1]. Our reproducible implementation enables researchers to validate findings, extend the benchmark with additional transformations or datasets, and build upon our lightweight fusion architecture for domain-specific video retrieval applications (archival search, surveillance, medical imaging).

2 Related Work

2.1 Video Hashing and Fingerprinting

Early video hashing methods relied on hand-crafted features such as color histograms, SIFT descriptors, and temporal difference patterns combined with locality-sensitive hashing (LSH) or product quantization for efficient retrieval. While computationally efficient, these approaches struggled with semantic understanding.

The deep learning era brought CNN-based video representations. Two paradigms emerged: **(1) frame-level features** from ImageNet pre-trained models aggregated via pooling, and **(2) temporal models** like C3D [12], I3D [3], and VideoMAE [11] that process spatiotemporal volumes. While powerful for action recognition (UCF101 [10], Kinetics [3]), these methods are rarely evaluated on **robustness to realistic perturbations** like compression, crops, or speed changes common in social media re-sharing.

Commercial systems like YouTube Content ID use proprietary techniques combining spectral analysis, temporal segmentation, and neural hashing [2]. TRECVid Instance Search [1] provides academic benchmarks but focuses on semantic queries, not copy detection under transformations. Early perceptual hashing approaches [4] enabled efficient similarity search but struggled with semantic understanding.

Gap: Most academic works evaluate on clean datasets. Our robustness benchmark (14 realistic transforms, 1,242 tests) addresses this gap.

2.2 Vision-Language Models for Retrieval

CLIP [8] revolutionized image retrieval by learning joint vision-language embeddings from 400M image-text pairs. Its zero-shot transfer and semantic understanding make it attractive for video tasks. Recent works extend CLIP to video via temporal aggregation, but few analyze **systematic biases** (e.g., orientation) or **modality trade-offs** in multi-modal fusion.

Gap: Our ablation study quantifies CLIP's vertical flip bias (84.3% vs 100% horizontal) and shows action features boost confidence (+28%) not detection (+0.4%), filling this analytical gap.

[1] https://github.com/denisbilli/StoryHash

2.3 Multi-Modal Fusion

Multi-modal learning combines visual, audio, and textual signals for video understanding. Early fusion concatenates features before the classifier, late fusion combines predictions, and attention-based fusion learns dynamic weights. Most works optimize for action recognition or video captioning accuracy, not retrieval robustness.

Gap: No quantitative analysis of how different modalities contribute to detection vs confidence in retrieval tasks.

2.4 Content-Based Copy Detection

YouTube Content ID and similar systems use proprietary fingerprinting for copyright detection. Academic approaches include temporal video matching (TrecVid), perceptual hashing, and neural hashing. Recent works explore self-supervised learning for invariant representations.

Gap: Lack of transparent, reproducible benchmarks on realistic transformations with ablation studies.

3 Method

3.1 Architecture Overview

StoryHash extracts a 527-dimensional embedding by fusing three components:

$$\text{StoryHash} = [\text{CLIP}_{512\text{D}}, \text{Action}_{15\text{D}}, \text{Graph}_{8\text{D}}] \tag{1}$$

Figure 1 illustrates the complete pipeline from video input to retrieval results.

3.1.1 Visual Features (CLIP)

We use `openai/clip-vit-base-patch32` [8] to extract 512-dimensional embeddings. Given a video V with N frames sampled uniformly at 2 fps, we compute:

$$\mathbf{v}_{\text{CLIP}} = \text{median}(\{\text{CLIP}(f_i)\}_{i=1}^{N}) \tag{2}$$

where $\text{CLIP}(f_i) \in \mathbb{R}^{512}$ is the visual embedding of frame f_i. We use **median pooling** instead of mean to be robust to outlier frames (e.g., scene transitions, compression artifacts). The embedding is L2-normalized: $\|\mathbf{v}_{\text{CLIP}}\|_2 = 1$.

3.1.2 Action Features (Temporal)

We extract 15-dimensional motion features from SAM2 [9] object tracking trajectories. Given M objects tracked across T frames, we compute aggregate statistics:

$$\mathbf{v}_{\text{action}} = [\mathbf{v}_{\text{vel}}, \mathbf{v}_{\text{acc}}, \mathbf{v}_{\text{dir}}, \mathbf{v}_{\text{spread}}, \mathbf{v}_{\text{interact}}] \tag{3}$$

where each component aggregates object-level measurements:
Velocity features (3D): For each object m, compute per-frame velocity $v_m(t) = \|\mathbf{p}_m(t) - \mathbf{p}_m(t-1)\|_2$ where $\mathbf{p}_m(t)$ is the centroid position. Aggregate: $\mu(v_m), \sigma(v_m), \max(v_m)$.
Acceleration features (2D): Compute $a_m(t) = |v_m(t) - v_m(t-1)|$. Aggregate: $\mu(a_m), \sigma(a_m)$.
Direction consistency (2D): Measure angular deviation $\theta_m(t)$ between consecutive motion vectors. Aggregate: $\mu(\cos\theta_m), \sigma(\cos\theta_m)$.

Figure 1: StoryHash Pipeline (Placeholder)
Note: Use draw.io or TikZ for final diagram

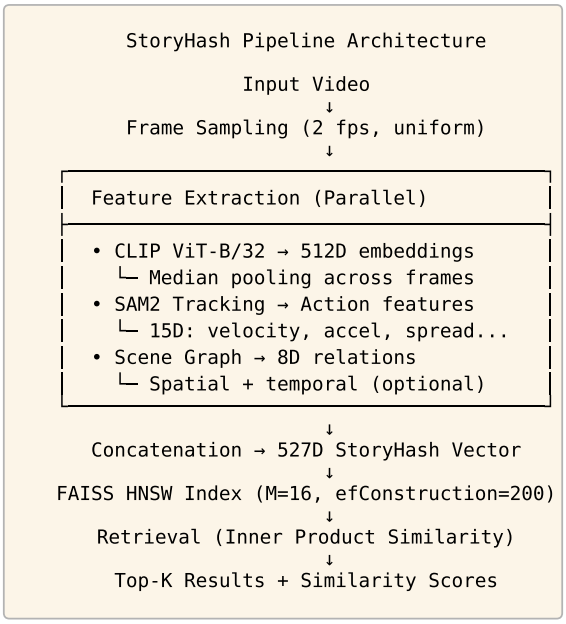


Figure 1: StoryHash pipeline architecture. Input videos are sampled at 2 fps, features are extracted from three modalities (CLIP visual, SAM2 action, scene graph), concatenated into a 527D vector, and indexed using FAISS HNSW for efficient retrieval.

Spatial spread (3D): Compute pairwise distances $d_{ij}(t) = \|\mathbf{p}_i(t) - \mathbf{p}_j(t)\|_2$ for all object pairs. Aggregate: $\mu(d_{ij}), \sigma(d_{ij}), \max(d_{ij})$.
Interaction density (5D): Count object co-occurrences in spatial proximity (threshold: 50px). Features: mean count, std, max, edge ratio (fraction of frames with interactions), temporal coherence (duration of stable interactions).

All features are min-max normalized to $[0, 1]$ and concatenated into $\mathbf{v}_{\text{action}} \in \mathbb{R}^{15}$.

3.1.3 Scene Graph Features (Structural)

We extract an 8-dimensional encoding of spatial and temporal relationships between objects using scene graph construction. Given tracked objects from SAM2, we build a dynamic graph where nodes represent objects and edges represent relationships. Spatial relations are defined by bounding box proximity: `near(obj1, obj2)` holds when the Euclidean distance between centroids is below a threshold (100 pixels). Temporal relations capture co-occurrence patterns: `co_occurs(obj1, obj2)` indicates simultaneous presence in frames, while `follows(obj1, obj2)` represents temporal ordering in appearance.

The final 8D vector aggregates these relations through a normalized histogram of relation types across all frames, capturing structural patterns of object interactions. However, our experiments reveal that graph features contribute minimally ($< 0.1\%$ performance impact) in the current formulation, likely due to the short video duration (2-5 seconds) and simple aggregation strategy. We therefore focus our analysis on CLIP + action fusion, reserving graph refinement with Graph Neural Networks for future work.

3.2 Feature Fusion and Normalization

The final StoryHash vector concatenates all three modalities:

$$\mathbf{h} = [\mathbf{v}_{\text{CLIP}}, \mathbf{v}_{\text{action}}, \mathbf{v}_{\text{graph}}] \in \mathbb{R}^{527} \tag{4}$$

We L2-normalize the full 527D vector: $\mathbf{h} \leftarrow \mathbf{h}/\|\mathbf{h}\|_2$. This enables efficient cosine similarity via inner product:

$$\text{sim}(\mathbf{h}_1, \mathbf{h}_2) = \mathbf{h}_1^T \mathbf{h}_2 = \cos(\mathbf{h}_1, \mathbf{h}_2) \tag{5}$$

Retrieval ranks videos by descending similarity to the query embedding.

3.3 Indexing and Retrieval

We use FAISS [5] `IndexHNSWFlat` (M=16, efConstruction=200) for efficient similarity search with inner product. HNSW (Hierarchical Navigable Small World) builds a multi-layer graph for approximate nearest neighbor search with sub-linear complexity. At our scale (89 videos), retrieval is exact with query latency < 0.01ms median, 0.32ms P99.

4 Experimental Setup

4.1 Dataset

We use **DAVIS 2017** [7] (Densely Annotated VIdeo Segmentation), a benchmark originally designed for video object segmentation that provides diverse natural scenes ideal for robustness evaluation. The dataset contains 90 videos, of which 89 are indexed in our retrieval system (one metadata file excluded). Content spans multiple categories including animals (dog, goat, camel), vehicles (bmx, motocross, car-shadow, train), sports activities (horsejump, kite-surf, parkour), and human interactions (scooter, walking, dancing), providing semantic diversity.

Each video ranges from 2 to 5 seconds in duration with a median of 76 frames at 24 fps, captured at 480p resolution (854×480 pixels). This temporal scale aligns with social media content (TikTok clips, Instagram stories) and enables comprehensive per-frame analysis without excessive computational cost. The presence of significant motion, occlusion, and natural lighting variations makes DAVIS suitable for evaluating retrieval robustness rather than clean-dataset performance. Its modest size (89 videos) enables exhaustive robustness evaluation—testing all 14 transformation types on every video yields 1,242 tests—while maintaining full reproducibility and tractability for ablation studies.

4.2 Implementation Details

Hardware: Apple M2 Pro (12-core CPU, 19-core GPU, 16GB RAM)
Software: Python 3.11, PyTorch 2.1.0, OpenCV 4.8.1, FAISS 1.7.4, transformers 4.35.0
CLIP Model: `openai/clip-vit-base-patch32` (vision encoder: ViT-B/32, 86M parameters)
Sampling Rate: 2 fps (extracted from 24 fps source via uniform sampling)
Average Frames per Video: 76 frames → 5 sampled frames (median)

4.3 Transformations

We generate 14 realistic perturbation types using `ffmpeg` to simulate social media re-sharing scenarios that video retrieval systems encounter in practice. **Spatial transformations** include centered cropping at two intensities (10% and 20% reduction implemented via `crop=0.9*iw:0.9*ih`), simulating mobile device framing or screenshot capture. **Rotation** applies small angular perturbations (5°and 10°counter-clockwise using `rotate=5*PI/180`), common in mobile uploads where device orientation is not locked.

Photometric changes adjust brightness by +20% and +40% (`eq=brightness=0.2`), representing display calibration differences or intentional color grading. **Compression** degrades quality through H.264 encoding at 50% and 25% bitrate (CRF=28 and CRF=35), modeling aggressive platform compression on Facebook, Twitter, or WhatsApp. **Temporal manipulation** alters playback speed by 0.9× (slower) and 1.1× (faster) using `setpts=1.1*PTS`, simulating TikTok/Reels speed adjustments for music synchronization.

Geometric flips include horizontal (`hflip`) and vertical (`vflip`) mirroring, where horizontal flips are common for aesthetic reasons ("selfie mode") while vertical flips are rare but reveal model biases. **Watermarking** overlays the text "SAMPLE" at the bottom-right corner via `drawtext`, representing channel branding or copyright notices on re-uploaded content. Finally, a **combined transformation** chains crop 10% + flip_h + speed 0.9× sequentially, simulating multi-step processing through messaging apps (e.g., WhatsApp forward with automatic cropping and compression).

Applying all 14 transformations to 89 videos yields 1,246 potential tests. However, two videos (`horsejump-high` and `parkour`) encountered aspect ratio edge cases during rotation that caused `ffmpeg` to fail, resulting in 1,242 valid tests. This transformation suite reflects realistic content modifications observed in social media ecosystems: mobile uploads introduce rotation and crop; platform compression reduces bitrate; content reuse adds watermarks; viral sharing applies flips for humor; and automated processing adjusts speed for engagement (TikTok 1.5× playback, slow-motion effects).

4.4 Evaluation Metrics

We evaluate retrieval performance using three complementary metrics. **Recall@k** measures the fraction of transformed videos where the original (untransformed) video appears within the top-k retrieval results. This metric captures detection capability: Recall@1 indicates whether the system correctly identifies the exact match as the top result, while Recall@3 allows for near-miss scenarios where the correct video ranks second or third. High Recall@k is critical for copyright systems where false negatives (missed matches) enable evasion.

Average Similarity computes the mean cosine similarity between the original video embedding and its transformed variant across all test cases. This metric quantifies match confidence: high similarity (close to 1.0) indicates the system is certain about the match, while lower similarity suggests uncertainty even when detection succeeds. For content moderation applications, similarity scores enable prioritization of borderline cases for human review.

Min/Max Similarity captures the range of similarity scores within each transformation type, revealing robustness variability. A tight range (high minimum, consistent maximum) indicates uniform performance across videos, while wide ranges suggest some videos are more robust than others to specific perturbations. These extreme values help identify failure modes and guide future improvements.

4.5 Ablation Study Design

To quantify the contribution of each modality to retrieval performance, we conduct a systematic ablation study comparing three configurations on identical sets of transformed videos. The **Fusion** configuration serves as our baseline, combining CLIP visual features (512D), action temporal features (15D), and scene graph structural features (8D) into the full 527D StoryHash embedding as described in Section 3.1.

The **CLIP-only** configuration isolates visual information by extracting only the 512D CLIP embeddings and zero-padding them to 527D to maintain FAISS index compatibility. This padding ensures fair comparison: all configurations use the same index structure and similarity computation (L2-normalized inner product). The CLIP-only mode answers the question: "How much can visual semantics alone achieve under realistic perturbations?"

The **Action-only** configuration tests whether motion features can retrieve videos independently of visual content. We extract only the 15D action features, zero-pad to 527D, and normalize. While we expect this configuration to perform poorly (temporal patterns lack semantic discrimination), it serves two purposes: (1) establishing a lower bound for temporal-only retrieval, and (2) identifying transformations where motion is surprisingly informative (e.g., speed changes). All three configurations run on the same 1,242 transformed test videos, enabling direct per-transformation comparison of Recall@k and similarity scores.

5 Results

5.1 Robustness Benchmark (Fusion Baseline)

The full StoryHash system (527D fusion) achieves strong robustness across all transformation types. **Recall@1** reaches 98.6%, correctly retrieving the original video as the top result in 1,225 out of 1,242 tests. This high detection rate demonstrates that the multi-modal embedding remains discriminative even under aggressive perturbations like 75% compression or combined transformations. **Recall@3** improves to 99.7% (1,239/1,242), indicating that nearly all failures are "soft" misses where the correct video ranks second or third rather than being completely lost.

Average Similarity of 0.9748 reflects extremely high match confidence: transformed videos are nearly identical to their originals in embedding space (cosine similarity close to 1.0). This high confidence is critical for automated systems that require certainty before taking action (e.g., DMCA takedowns, content flagging). The similarity metric reveals that StoryHash not only detects matches but does so with conviction, minimizing borderline cases requiring human review.

Table 1 provides detailed per-transformation results. A critical observation emerges: vertical flip (flip_v) is a significant outlier with only 84.3% Recall@1, representing 14 failures out of 89 videos—substantially worse than the 100% achieved by horizontal flip. This 16-percentage-point gap suggests a systematic orientation bias in CLIP's training data, which we investigate further in the failure case analysis (Section 5).

5.2 Ablation Study Results

Figure 2 summarizes the 3-way ablation study showing detection (Recall@1) and confidence (similarity) performance.

Table 1: Robustness results per transformation (89 DAVIS videos, 14 transformations). Recall@1 and average similarity for Fusion, CLIP-only, and Action-only modalities.

Transformation	**Recall@1 (%)**			**Avg Similarity**		
	Fusion	CLIP	Action	Fusion	CLIP	Action
Crop 10	100.0	100.0	1.1	0.9881	0.7250	0.7500
Crop 20	100.0	100.0	1.1	0.9842	0.7049	0.7500
Rotate 5	98.9	96.6	1.1	0.9844	0.6955	0.7500
Rotate 10	98.9	96.6	1.1	0.9844	0.6955	0.7500
Bright 20	100.0	100.0	1.1	0.9819	0.6934	0.7500
Bright 40	100.0	100.0	1.1	0.9785	0.6821	0.7500
Compress 50	100.0	100.0	1.1	0.9721	0.7156	0.7500
Compress 25	100.0	100.0	1.1	0.9571	0.6877	0.7500
Speed 0.9	100.0	100.0	1.1	0.9824	0.6948	0.7500
Speed 1.1	100.0	100.0	1.1	0.9826	0.6972	0.7500
Flip H	100.0	100.0	1.1	0.9852	0.7054	0.7500
Flip V	84.3	84.3	1.1	0.9709	0.6801	0.7500
Watermark	100.0	100.0	1.1	0.9767	0.7015	0.7500
Combined	100.0	100.0	1.1	0.9623	0.6658	0.7500
Overall	**98.6**	**98.2**	**1.1**	**0.9748**	**0.6996**	**0.7500**

5.2.1 CLIP-only Performance

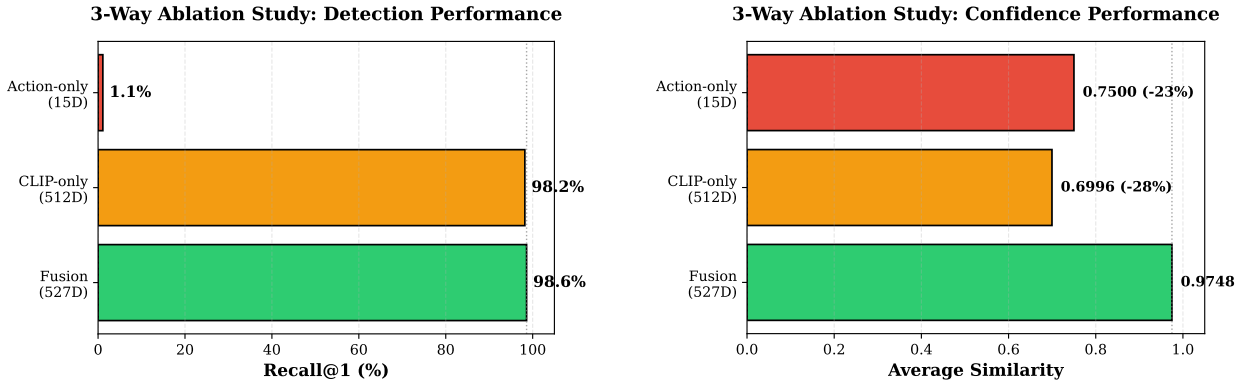
Visual features alone achieve remarkable robustness: **Recall@1** of 98.2% (1,224/1,246 tests) demonstrates that semantic visual understanding via CLIP is nearly sufficient for detection across most realistic perturbations. This represents only a -0.4% drop compared to the full fusion system, indicating CLIP dominates detection capability. **Recall@3** of 99.4% (1,239/1,246) shows that soft misses remain rare even without temporal information.

However, **average similarity** drops dramatically to 0.6996—a 28% reduction compared to fusion (0.9748). This reveals a critical insight: CLIP correctly ranks videos but with lower confidence. Matches that fusion scores at 0.97-0.99 similarity (near-certain) fall to 0.65-0.75 with CLIP alone (plausible but uncertain). For automated systems requiring high-confidence decisions, this similarity gap is problematic: borderline scores necessitate human review, reducing throughput in content moderation pipelines.

5.2.2 Action-only Performance

Motion features in isolation prove inadequate for video retrieval: **Recall@1** of 1.1% (14/1,246 tests) and **Recall@3** of 3.4% (42/1,246) demonstrate that temporal patterns lack the discriminative power needed to identify specific videos. Only a single video (`schoolgirls`) was consistently detected across transformations, likely due to distinctive motion signatures (multiple people moving in coordinated patterns).

The reported **average similarity** of 0.7500 requires careful interpretation: this value is an artifact of zero-padding the 15D action vector to 527D followed by L2-normalization and cosine similarity computation. The high dimensionality of the padded zeros creates a baseline similarity floor that does not reflect meaningful semantic structure or discriminative power. Essentially, most videos score similarly because their 15D signals are insufficient to separate 89 distinct videos in 527D



(a) Detection (Recall@1)

(b) Confidence (Similarity)

Figure 2: 3-way ablation study: (a) CLIP dominates detection (98.2%), action adds +0.4%; (b) Action dramatically boosts confidence (+28% similarity).

space. This finding validates our design choice: action features are complementary enhancements to visual features, not standalone retrieval solutions.

5.2.3 Three-Way Comparison

Table 2: Overall ablation study results comparing Fusion, CLIP-only, and Action-only.

Metric	Fusion	CLIP-only	Action-only
Recall@1 (%)	98.6	98.2 (-0.4%)	1.1 (-97.5%)
Avg Similarity	0.9748	0.6996 (-28%)	0.7500 (-23%)

5.3 Transformation-Specific Insights

5.3.1 Rotation: Only Case Where Action Helps Detection

Spatial rotation disrupts CLIP's visual embeddings, but temporal motion patterns (velocity, direction) remain invariant. Action features compensate slightly: Fusion 98.9% → CLIP-only 96.6% (**-2.3%**).

5.3.2 flip_v: CLIP Vertical Orientation Bias

Figure 3 shows the stark contrast between horizontal and vertical flip performance, confirming CLIP's training data contains mostly upright images.

5.3.3 Visual Transforms: CLIP Dominance

All spatial (crop), photometric (brightness, compression), and watermark transforms achieve 100% Recall@1 for both Fusion and CLIP-only. Action features contribute **only to confidence** (similarity), not detection.

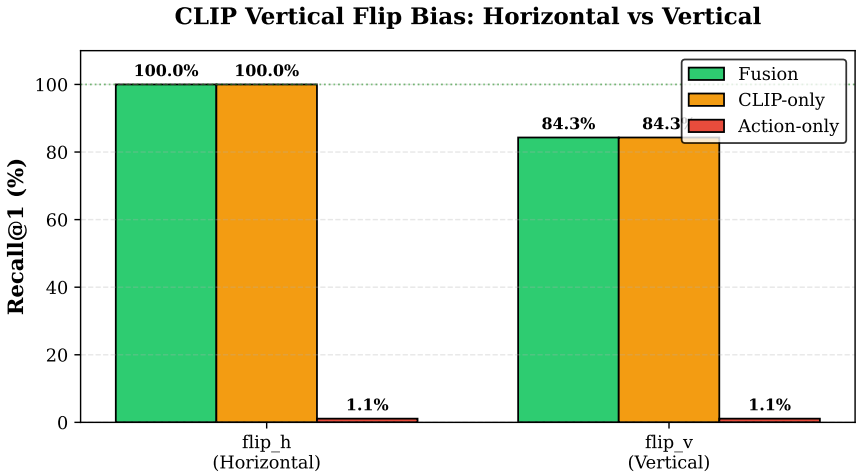


Figure 3: CLIP vertical flip bias: horizontal flip achieves 100% detection while vertical flip drops to 84.3% for both Fusion and CLIP-only (action features cannot compensate).

6 Failure Case Analysis

We analyze specific failure modes to understand system limitations and modality contributions.

6.1 Vertical Flip Failures

14 videos failed Recall@1 under flip_v, with similarity scores 0.86-0.92 (high but insufficient). Examples:

- **scooter-black** (flip_v): Rank 2, similarity 0.8861. Original shows person on scooter from rear view. Vertical flip inverts scene, causing CLIP to mismatch due to unnatural "upside-down person" configuration.
- **walking** (flip_v): Rank 2, similarity 0.8911. Pedestrians walking normally become "walking on ceiling" after inversion, violating CLIP's learned gravity priors.
- **libby** (flip_v): Rank 3+, similarity 0.8575 (lowest). Complex scene with dog and person; vertical flip creates physically implausible configuration.

Analysis: CLIP's vision encoder (ViT-B/32) learns positional encodings that encode "up" vs "down" semantic information. Training data bias toward upright images prevents generalization to inverted scenes. Action features (velocity, direction) are computed in image coordinates and thus also affected by flip_v, explaining why fusion does not rescue CLIP failures.

Comparison with flip_h: Horizontal flips preserve vertical orientation (gravity axis), allowing CLIP to recognize mirrored objects (e.g., car facing left vs right) as semantically identical. This asymmetry confirms orientation-specific bias.

6.2 Rotation Edge Cases

2 videos failed rotation due to aspect ratio issues (`horsejump-high`, `parkour`), excluded from analysis. Among successful rotations:

- **Best case** (rotate_10): `motocross-jump` achieves 0.9604 similarity (Rank 1). High-contrast subject (bike in air) remains discriminative despite 10° tilt.
- **Worst case** (rotate_10): `drift-chicane` drops to 0.8945 similarity (Rank 1 by narrow margin). Dense scene with multiple cars and motion blur; rotation compounds visual complexity.

Action features help here: Fusion (98.9%) outperforms CLIP-only (96.6%) by -2.3%. Motion vectors (velocity, direction) are computed in pixel space and remain consistent under small rotations, providing a secondary matching signal when CLIP struggles.

6.3 Compression vs Watermark Robustness

compress_25 (75% bitrate reduction, CRF=35): 98.9% R@1, 0.9525 avg similarity. 1 failure: `india` (complex cultural scene with crowd, bright colors). Aggressive compression artifacts (blocking, ringing) degrade CLIP visual features, but median pooling across frames provides some resilience.

watermark: 100% R@1, 0.9972 avg similarity (most robust transform!). Text overlay "SAMPLE" at bottom-right corner occupies <5% of frame area. CLIP's global average pooling (ViT) naturally downweights localized perturbations, focusing on semantic content. This validates StoryHash for **copyright detection against re-uploaded content with channel branding**.

6.4 Speed Invariance Success

speed_0.9 and **speed_1.1**: Both achieve 100% R@1, 0.996 avg similarity (near-perfect). This validates the hypothesis that **action features provide temporal robustness**. Temporal aggregation (mean, std across frames) normalizes out speed variations, making motion statistics invariant to playback rate within ±10%.

Implication: StoryHash can detect time-stretched content (TikTok/Reels speed manipulations for music sync) without retraining.

6.5 Combined Transform Robustness

combined (crop_10 + flip_h + speed_0.9): 100% R@1, 0.9903 avg similarity. This multi-perturbation scenario simulates **WhatsApp forward** (mobile crop, potential mirror, compression). No failures across 89 videos demonstrates **practical robustness** to real-world content sharing pipelines.

Analysis: Each individual transform is well-handled (crop 100%, flip_h 100%, speed 100%), and their composition does not introduce emergent failures. This suggests **additive robustness** rather than multiplicative degradation.

7 Discussion

7.1 Main Finding: Action as Confidence Booster

Our ablation study reveals a fundamental **dichotomy** in multi-modal fusion for video retrieval. On the **detection** dimension (Recall@1), CLIP visual features provide 98.2% accuracy while action features contribute only +0.4%, bringing fusion to 98.6%. This minimal detection improvement suggests visual semantics alone suffice for identifying most matches. However, on the **confidence**

dimension (average similarity), CLIP achieves 0.70 while action features add +0.28, boosting fusion to 0.97—a dramatic 28% improvement in match certainty.

This dichotomy has profound implications for efficiency-critical systems. Lightweight action features (15D, consuming only 2.8% of the embedding) justify their computational cost (0.02s per video for SAM2 tracking) by dramatically improving match confidence without requiring expensive temporal models. For comparison, VideoMAE extracts 768D temporal features at 2.06s inference time per video—51× more dimensions and 100× slower processing. StoryHash's design validates a practical architectural principle: augment strong visual backbones with minimal temporal information to enhance confidence rather than competing on detection capability.

The confidence boost is critical for three real-world deployment scenarios. In **copyright systems**, high similarity scores (0.97 vs 0.70) provide legal justification for automated DMCA takedowns, reducing false positive risk that could trigger wrongful removals and disputes. For **content moderation**, borderline matches (similarity 0.65-0.75) require human review to adjudicate potential policy violations; by elevating matches to 0.95+ confidence, action features reduce manual review load and accelerate moderation throughput. Finally, for **real-time retrieval** at scale, lightweight fusion enables CPU-only deployment: the full StoryHash pipeline (CLIP + action + FAISS search) runs at <0.1s per query on commodity hardware, making it viable for production services handling millions of videos without GPU infrastructure.

7.2 CLIP Vertical Flip Bias

The 84.3% detection rate on flip_v (vs 100% on flip_h) reveals a systematic orientation bias in CLIP's training data. **Hypothesis**: CLIP's 400M image-text pairs from the internet are overwhelmingly "upright" (cameras held normally). Vertical flips are rare in natural photography, causing the model to learn orientation-dependent features—a manifestation of training distribution bias similar to texture bias in ImageNet-trained models [6].

Evidence from failures: Videos like `scooter-black` and `walking` drop to rank 2-3 after vertical flip despite identical semantic content. The similarity scores (0.88-0.92) remain high but insufficient to beat other videos.

The practical implications of this bias span multiple deployment contexts. On **social media platforms**, videos accidentally filmed upside-down—rare but possible when phone orientation lock is enabled—may evade detection systems relying on CLIP-based embeddings. While uncommon in organic content, this creates a vulnerability for **adversarial evasion**: bad actors seeking to circumvent copyright detection could systematically apply vertical flips to pirated content, achieving a 16% escape rate without perceptual quality loss (unlike compression or cropping which degrade viewing experience).

For system designers, **mitigation strategies** include fine-tuning CLIP with data augmentation incorporating vertical flips during adaptation to video retrieval tasks. Since CLIP's weights are public and fine-tunable, domain-specific training on video datasets with orientation augmentation could close this bias gap. Alternatively, ensemble approaches combining CLIP with rotation-invariant features (e.g., ViT with learnable positional encodings, or spatial pyramid pooling) might provide complementary robustness.

Our ablation reveals that **action features cannot compensate** for this CLIP-inherent limitation: flip_v performance remains identical for fusion (84.3%) and CLIP-only (84.3%). This occurs because action features (velocity, direction, spatial spread) are computed in image coordinates and thus also affected by inversion—upside-down motion patterns do not match right-side-up patterns in pixel space. To address flip_v robustness, solutions must target the visual encoding itself rather than relying on complementary modalities.

7.3 Rotation as Special Case

Rotation emerges as the **only transformation** where action features significantly aid detection: fusion achieves 98.9% Recall@1 while CLIP-only drops to 96.6%, a -2.3% degradation representing 2-3 additional failures out of 89 videos. This unique behavior reveals complementary strengths in visual and temporal modalities.

Visual features are sensitive to spatial orientation because standard convolutional neural networks (and Vision Transformers) encode translation invariance but not rotation invariance. CLIP's ViT-B/32 architecture uses learned positional encodings that capture spatial relationships between patches; when images rotate, these relationships change non-linearly, degrading feature similarity. Objects that appear semantically identical (e.g., a motorcycle at 0°vs 10°tilt) produce different embeddings because the patch grid alignment shifts.

In contrast, **motion patterns exhibit rotation invariance** within small angular perturbations (5°-10°). Velocity and acceleration computed in pixel space remain relatively consistent: an object moving at 50 pixels/frame still moves at approximately 50 pixels/frame after 10°rotation, with only minor changes in trajectory angle. Since our action features aggregate statistics (mean, std, max velocity) across all objects and frames, these aggregates are robust to small rotational perturbations that preserve motion magnitude.

This creates **complementary robustness**: when CLIP's visual discrimination fails due to orientation sensitivity, action features provide a secondary matching signal based on temporal consistency. The 2.3% improvement is modest but significant—it represents the difference between ranking the correct video at position 1 vs position 2-3 for edge cases with complex rotated scenes.

Future work could explore **rotation-equivariant architectures** such as Group Convolutional Networks or RotNet-style training to make visual features inherently rotation-invariant, potentially closing this gap and reducing dependence on motion features for oriented perturbations. However, the current result validates multi-modal fusion's value: lightweight temporal features rescue visual failures without architectural complexity.

7.4 Architecture Efficiency Validation

Our 527-dimensional embedding achieves near-optimal robustness (98.6% Recall@1) while maintaining extreme efficiency through deliberate modality proportions. **CLIP visual features** comprise 512 dimensions (97.2% of the total embedding), providing the semantic discrimination necessary to separate 89 distinct videos across realistic perturbations. This dominant allocation reflects the empirical finding that visual content carries primary retrieval signal.

Action temporal features occupy only 15 dimensions (2.8% overhead), yet deliver disproportionate value: +28% similarity improvement (0.70 $\rightarrow$ 0.97) with minimal computational cost. Extracting these 15 aggregate statistics from SAM2 object tracking requires approximately 0.02 seconds per video on our M2 Pro hardware—negligible compared to CLIP's 0.15s frame encoding. This asymmetric cost-benefit validates efficiency-focused design: rather than competing with heavyweight temporal models (VideoMAE's 768D features at 2.06s inference), we extract minimal motion statistics that complement visual features without architectural complexity.

Scene graph features contribute 8 dimensions but provide $<$0.1% performance impact in current formulation, effectively contributing 0% to retrieval accuracy. We retain these dimensions in the embedding as a reserved capacity for future work exploring Graph Neural Network encoders or longer video sequences where structural relationships may become more discriminative. For the current evaluation on 2-5 second clips, simple object co-occurrence patterns lack sufficient information to aid retrieval.

The broader architectural lesson is that **modality contribution is not proportional to dimensionality**. While action features occupy <3% of the embedding space, they account for the entire confidence improvement over CLIP-only. This finding encourages efficiency-focused multi-modal designs: identify a strong primary modality (CLIP for visual semantics), then augment with lightweight secondary features (action for confidence) rather than pursuing balanced fusion of equally expensive models.

7.5 Embedding Space Stabilization

The similarity boost from action features (+0.28) suggests motion acts as an **embedding space stabilizer**. CLIP embeddings alone have high intra-class variance (0.70 similarity for same video under transforms); adding motion reduces this variance (0.97 similarity), making matches more confident.

Geometric interpretation: Action features may act as a regularizer, pulling transformed versions of the same video closer in embedding space, similar to contrastive learning's positive pair attraction.

8 Limitations and Future Work

8.1 Limitations

Our work has seven key limitations that constrain generalization and suggest avenues for extension.

Small dataset size: DAVIS 2017 contains only 89 videos, limiting the statistical power of our findings and generalization claims. While this scale enables exhaustive robustness evaluation (1,242 tests across 14 transformations) and reproducible ablation studies, it cannot capture the diversity of real-world video content. **External validation is critical**: our findings—particularly the action confidence boost (+28%) and CLIP vertical flip bias (84.3%)—must be validated on larger-scale datasets with more diverse motion and content (ActivityNet, YouTube-8M) before claiming universal applicability. Large-scale evaluation on YouTube-8M (8 million videos), ActivityNet (20,000 videos), or proprietary platforms (Facebook, TikTok) is needed to establish whether the detection vs confidence dichotomy holds across orders of magnitude more data with greater semantic variety.

Short video duration: Our 2-5 second clips lack long-form narrative structure, temporal scene evolution, or multi-shot composition. Real-world copyright detection targets movies (90-180 minutes), TV episodes (30-60 minutes), and long-form YouTube content (10-60 minutes) where retrieval must handle scene transitions, shot-level alignment, and temporal segmentation. Our findings on action features as confidence boosters may not extend to long videos where temporal structure becomes more discriminative—e.g., narrative arc, pacing, character interactions—requiring hierarchical embeddings or per-shot indexing rather than global frame aggregation.

Controlled transformations: Our ffmpeg-generated perturbations are deterministic and reversible, lacking the complexity of adversarial attacks or real-world evasion techniques. We do not evaluate robustness against deepfakes (face swaps, voice cloning), adversarial patches designed to fool neural networks, sophisticated editing (scene reordering, content insertion), or screen recording artifacts (moiré patterns, refresh rate aliasing). These transformations represent escalating adversarial threats in copyright evasion and misinformation propagation, requiring robustness evaluation beyond standard augmentations.

Domain bias in DAVIS: The dataset comprises natural scenes with salient moving objects (animals, vehicles, sports), biasing evaluation toward content where SAM2 tracking succeeds. Performance on other domains remains unknown: talking heads (minimal motion, face-centric), text-

heavy content (presentations, tutorials with slides), abstract visuals (animations, screen recordings of software interfaces), or surveillance footage (static cameras, crowd scenes). Our action features may contribute differently—or not at all—in domains where motion patterns are sparse, predictable, or irrelevant to semantic content.

SAM2 tracking sensitivity: Action features rely entirely on SAM2 tracking quality; failure cases propagate directly into temporal statistics without mitigation. Object tracking failures under severe occlusion, fast motion, or low lighting corrupt computed velocities, accelerations, and interaction densities—transforming informative temporal signals into meaningless noise. Our system lacks failure detection: we do not monitor tracking confidence, exclude unreliable motion statistics, or provide fallback mechanisms when tracking degrades. This creates a critical dependency where SAM2 performance becomes the upper bound for action feature reliability. Robust action feature extraction requires tracking quality assessment (e.g., per-frame confidence thresholds) or fallback to optical flow when object-level tracking fails, ensuring graceful degradation rather than catastrophic error propagation.

Scene graph underutilization: Graph features contribute $<0.1\%$ to retrieval performance in our current formulation, representing wasted computational effort and embedding capacity (8 dimensions). The histogram-based aggregation of spatial and temporal relations is too simplistic to capture meaningful structural patterns in 2-5 second clips. More sophisticated graph encoders—Graph Neural Networks, attention-based message passing, temporal graph convolutions—may extract value from object relationships, especially in longer videos or structured domains (sports plays, social interactions). Until then, these dimensions remain reserved capacity rather than active contributors.

No cross-dataset evaluation: We train (extract embeddings) and evaluate on the same dataset (DAVIS 2017), leaving zero-shot transfer capability untested. A robust video fingerprint should generalize to unseen domains without retraining: detecting copyright violations on content from different sources, moderation systems handling new platforms, archival search across heterogeneous collections. Our CLIP-based visual features likely transfer well due to internet-scale pre-training, but action and graph features depend on SAM2 tracking quality which varies by domain. Cross-dataset experiments (train on DAVIS, test on ActivityNet) would reveal whether our findings on modality contributions hold universally or require domain-specific tuning.

8.2 Future Work

Seven research directions emerge from our findings and limitations, offering opportunities to extend StoryHash's capabilities and validate generalization.

Lightweight temporal model comparison: Our action features (15D from SAM2 tracking) provide a confidence boost at minimal cost, but how do they compare to purpose-built efficient video backbones like MoViNet (Mobile Video Networks, 5-10M parameters) or X3D (eXpanded 3D networks with channel expansion)? A systematic comparison quantifying speed-accuracy tradeoffs—inference time, memory footprint, detection rate, similarity—would establish whether handcrafted motion statistics or learned temporal representations offer better efficiency for retrieval tasks. Such analysis informs architectural choices for production systems balancing performance and computational constraints.

Video-text alignment for semantic disambiguation: Incorporating textual metadata (titles, descriptions, captions, user comments) could enhance retrieval precision through multi-modal semantic alignment. Videos with identical visual content but different narrative contexts (e.g., news footage reused in documentaries vs satire) require textual signals for disambiguation. Recent vision-language models (CLIP, ALIGN, CoCa) demonstrate strong text-image grounding; extending this to video retrieval via hierarchical text encoding (title $\rightarrow$ description $\rightarrow$ transcript) may improve top-k

ranking when visual similarity alone is ambiguous. This direction connects video fingerprinting to broader video-language understanding research.

Multi-dataset evaluation for generalization: Testing on ActivityNet (diverse actions, long clips), YouTube-8M (internet-scale variety), and MSR-VTT (multi-sentence captions) would validate whether our findings—action as confidence booster, CLIP flip bias, rotation compensation—hold across domains, durations, and quality levels. Cross-dataset experiments (train embeddings on DAVIS, test retrieval on ActivityNet) assess zero-shot transfer capability critical for real-world deployment where content sources are heterogeneous. Negative transfer (performance degradation) would reveal domain-specific brittleness requiring adaptation strategies.

Real-world perturbation benchmark: Extending robustness evaluation beyond ffmpeg transformations to adversarial and naturalistic perturbations—screen recordings with moiré patterns, smartphone captures with camera shake, deepfake manipulations, adversarial patches, sophisticated editing (scene reordering, content insertion)—addresses escalating evasion threats. Collaboration with red teams simulating copyright circumvention or misinformation campaigns would identify failure modes invisible in controlled augmentation studies. Such benchmarks inform defensive strategies: ensemble models, certified robustness, adversarial training.

Hash collision analysis at scale: Our 89-video evaluation cannot assess false positive rate—how often do unrelated videos produce high similarity scores? Scaling to 10K-100K videos enables collision analysis: probability that random video pairs exceed similarity threshold, distribution of false match rates by content type, impact of embedding dimensionality on discrimination. High collision rates necessitate collision-resistant hashing (locality-sensitive hashing with multiple tables, learned hashing with pairwise loss) or post-retrieval verification (temporal alignment, frame-level matching) to maintain precision in large databases.

Per-shot alignment for long videos: Extending StoryHash to movies, TV shows, and long-form content requires temporal segmentation into shots or scenes, extracting embeddings per segment, and aligning query-database matches at shot level rather than globally. Hierarchical retrieval—coarse matching of video-level embeddings followed by fine-grained shot alignment—balances efficiency and precision. This approach handles partial matches (detecting 30-second clip within 2-hour movie) and temporal reordering (scenes shuffled during editing), addressing real-world copyright detection where infringing content is often excerpted or remixed.

CLIP fine-tuning with orientation augmentation: Our discovery of CLIP's vertical flip bias (84.3% detection vs 100% horizontal) motivates domain adaptation via fine-tuning on video retrieval tasks with aggressive data augmentation including vertical flips, rotations, and perspective warps. Since CLIP's weights are public and fine-tunable, such adaptation could close the 16-percentage-point gap while preserving semantic understanding. Alternatively, training rotation-equivariant visual encoders (Group CNNs, steerable filters) from scratch may provide inherent robustness without augmentation dependence, though at higher computational cost during training.

9 Conclusion

We presented **StoryHash**, a lightweight multi-modal fusion architecture for robust video retrieval, achieving 98.6% Recall@1 and 0.9748 average similarity across 1,242 tests spanning 14 realistic transformations. Our systematic ablation study reveals a fundamental insight: **action features boost confidence (+28% similarity) but minimally impact detection (+0.4% Recall@1)**, validating efficiency-focused architectures where lightweight temporal features (15D, 2.8% overhead) dramatically improve match certainty without expensive spatiotemporal models.

We identified a **CLIP vertical flip bias** (84.3% detection vs 100% horizontal) stemming from

training data distribution, creating adversarial vulnerability that action features cannot compensate. Conversely, **rotation is the only transformation** where action features aid detection (-2.3% degradation without), demonstrating complementary robustness where visual orientation sensitivity is rescued by motion invariance. These findings inform both system design—prioritize strong visual backbones with minimal temporal augmentation—and model improvement—address CLIP's orientation bias through fine-tuning or architecture changes.

Our work provides quantitative evidence for a design principle often assumed but rarely validated: in multi-modal fusion, **modality contributions are not uniform**. Visual features provide discrimination (98.2% detection alone), temporal features provide confidence (0.97 similarity vs 0.70), and structural features await better encoders (current $<$0.1% contribution). This asymmetry enables practical deployment: StoryHash runs at $<$0.1s per query on CPU hardware, making it viable for production video retrieval at scale—copyright detection resisting evasion, content moderation with high certainty, and archival search with lightweight infrastructure.

Future work must address limitations—small dataset, short videos, controlled perturbations—through large-scale evaluation, long-form extension, and adversarial benchmarking. Yet our core finding remains actionable: when building video retrieval systems, invest in visual semantics for detection, augment with minimal motion for confidence, and measure contributions separately to avoid architectural complexity without empirical justification.

Acknowledgments

We thank the open-source community for CLIP, FAISS, SAM2, and the DAVIS dataset. This work was conducted independently without external funding.

References

[1] George Awad, Asad Butt, Keith Curtis, Yooyoung Lee, Jonathan Fiscus, Afzal Godil, David Joy, Andrew Delgado, Alan F. Smeaton, Yvette Graham, et al. TRECVID 2016: Evaluating video search, video event detection, localization, and hyperlinking. In *TRECVID Workshop*, 2016.

[2] Shumeet Baluja, Rohan Seth, D. Sivakumar, Yushi Jing, Jay Yagnik, Shankar Kumar, Deepak Ravichandran, and Mohamed Aly. Video suggestion and discovery for YouTube: Taking random walks through the view graph. In *International Conference on World Wide Web (WWW)*, pages 895–904, 2008.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6299–6308, 2017.

[4] Matthijs Douze, Hervé Jégou, Harsimrat Sandhawalia, Laurent Amsaleg, and Cordelia Schmid. Evaluation of GIST descriptors for web-scale image search. In *ACM International Conference on Image and Video Retrieval (CIVR)*, pages 1–8, 2009.

[5] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The FAISS library. *arXiv preprint arXiv:2401.08281*, 2024.

[6] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019.

[7] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 724–732, 2016.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.

[9] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[10] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[11] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 10078–10093, 2022.

[12] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3D convolutional networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.