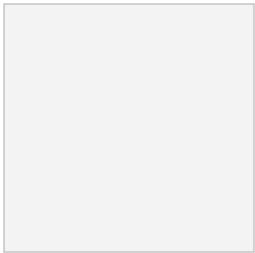


TinyShield

A browser extension for real-time detection
and blocking of AI-generated animated videos



Ingunn Fløvig



Sonal Khosla



Vaibhav Bajpai

ACM IMC 2026 Student Workshop | 12 October 2026

**Design IT.
Create Knowledge.**

www.hpi.de



Why now? Synthetic animation is becoming cheap to scale

A new supply problem meets a weak disclosure signal

01 Production

Text-to-video tools make it inexpensive to produce short, colorful animated clips in volume.

02 Exposure

Investigations found AI-generated children's videos in YouTube Shorts, and some on YouTube Kids.

03 Control gap

Clearly unrealistic animation may not require a YouTube AI disclosure. Absence of a label is not proof.

The question is not whether all AI animation is harmful — it is whether caregivers can choose, and researchers can measure.

TinyShield

Browser-side detection, blocking and measurement of synthetic children's animation

What if families could switch off AI-generated animated videos — without waiting for platform labels?

The New York Times

Fairplay's open letter, signed by 135 organizations and about 100 experts, **calls for a ban** on AI-generated children's content.

How A.I.-Generated Videos Are Distorting Your Child's YouTube Feed

Experts caution that low-quality, A.I.-generated videos on YouTube geared toward children often feature conflicting information, lack plot structure and can be cognitively overwhelming — all of which could affect young children's development.

NYT, "How A.I.-Generated Videos Are Distorting Your Child's YouTube Feed," 26 Feb 2026 · [nytimes.com/2026/02/26/us/ai-videos-children-youtube.html](https://www.nytimes.com/2026/02/26/us/ai-videos-children-youtube.html)

Regulation raises expectations, not a ready-made filter

Different rules address different parts of the problem

DSA · in force

Article 28 requires platforms accessible to minors to provide a high level of privacy, safety and security.

EU AI Act · Art. 50

Transparency rules for synthetic content apply from Aug 2026, with nuances for creative works and some transitional obligations.

EU KIDS Act · proposed

September 2026 proposal: child safety by design, controllable recommendation systems and age assurance. Not yet law.

None makes an unlabeled animated video automatically identifiable to an extension.

The measurement pipeline

RESEARCH QUESTIONS

RQ1 Can an extension detect and block while users browse?

PROTOTYPE BUILT

RQ2 How well can we detect synthetic children's animation?

BENCHMARK + SPECIALIZED DETECTOR PLANNED

RQ3 How often is it surfaced in recommendations?

RESEARCHER-OPERATED YOUTUBE / SHORTS / KIDS PROFILES
PLANNED

CONTRIBUTION AND SCOPE

BROWSER EXTENSION

Detect and block videos in the browser

PROTOTYPE BUILT

ANIMATION DETECTOR

Build and validate a detector for animation

PLANNED

MEASUREMENT STUDY

Estimate prevalence and repeated exposure

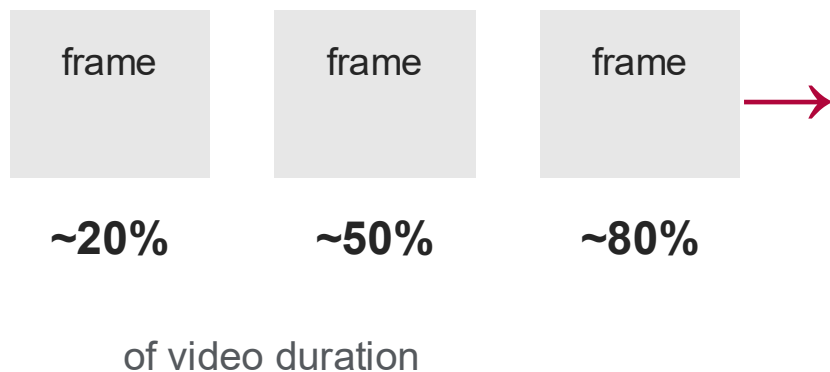
PLANNED

Measure both unique videos and repeated impressions — they answer different questions.

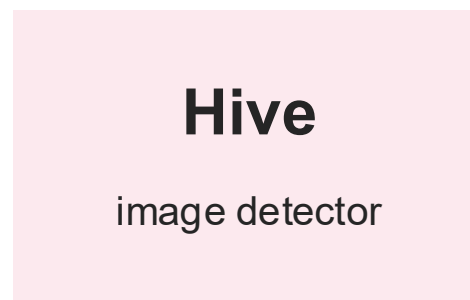
Three frames become one video-level decision

SAMPLE

3 frames per video

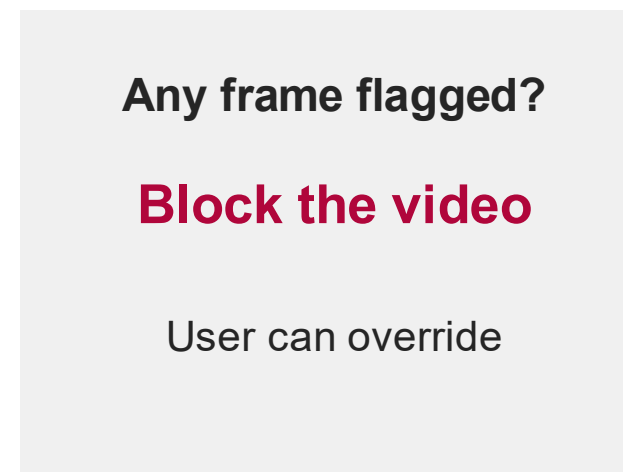


CLASSIFY



Replaceable module

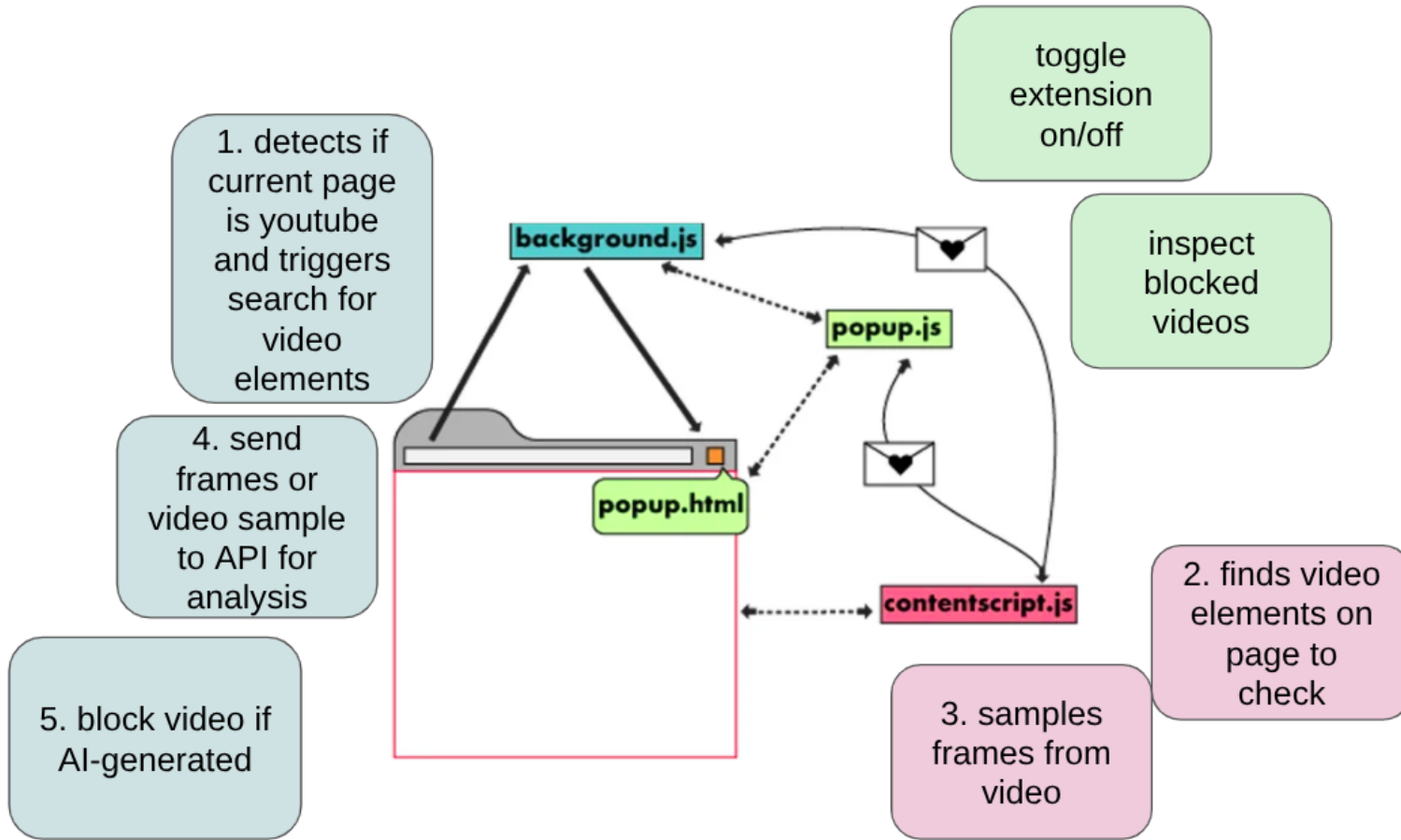
ACT



One false-positive frame can block an entire video.

Frame positions indicate sampling percentages; real-time latency remains to be evaluated.

TinyShield workflow (prototype)



Why Hive? Pragmatic integration, not proven accuracy

Detection is a modular component

Prototype choice: Hive

Available image-analysis API → integrate sampled frames into a working browser extension quickly.
Its accuracy on animated kids' videos is not yet validated.

Alternatives considered

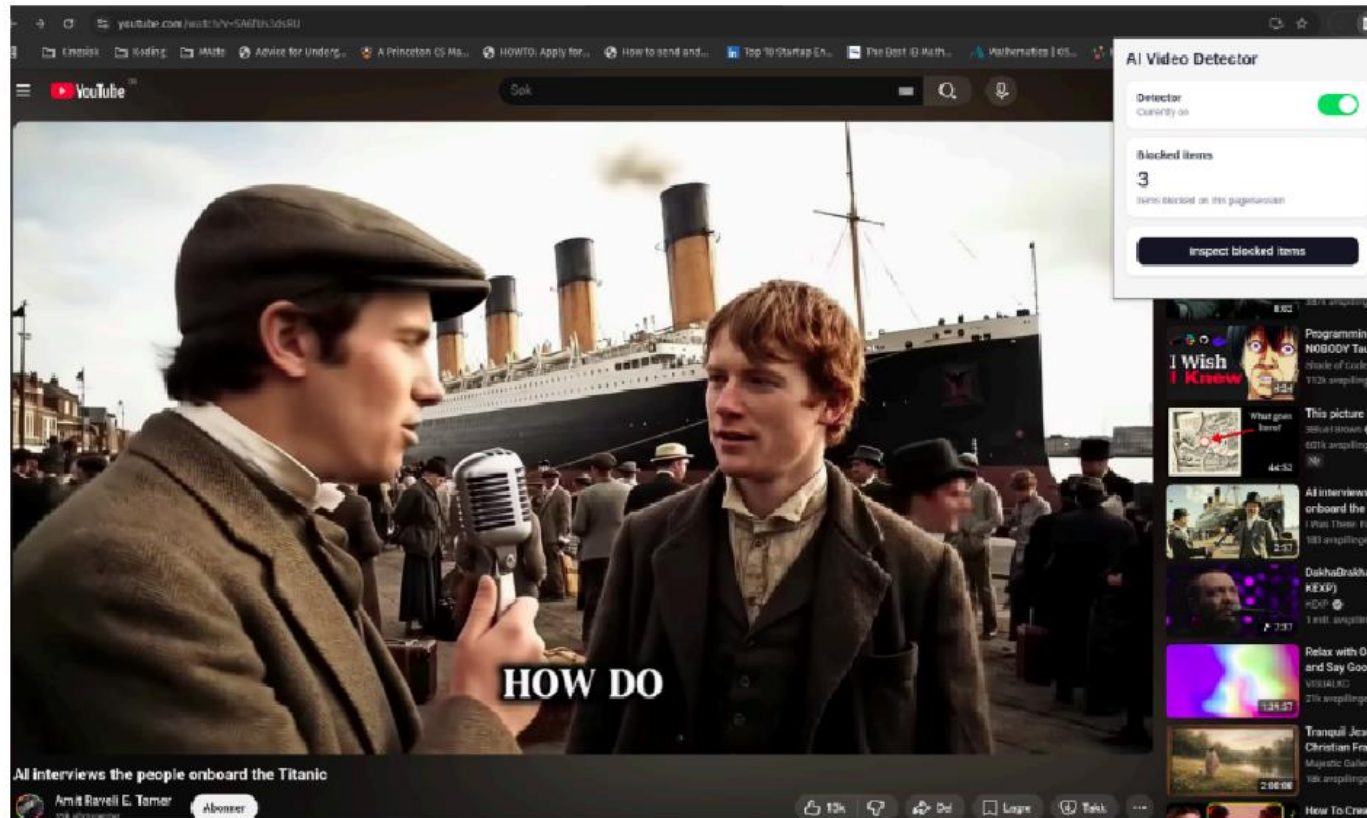
Commercial: Sightengine and other video APIs.
Research: Skyra, DeMamba, DeCoF; candidates include UNITE and ReStraV.

Pilot only: Skyra mislabeled 12/12 sampled videos in exploratory testing; no representative benchmark or error bars.

Sources: project detector survey (Mar 2026); Ingunn research draft (Jun 2026). Do not generalize the 12-video pilot.

The prototype works — with an important caveat

On/off toggle · blocked count · inspect and override



Parent choice

Toggle filtering on / off.

Explain and inspect

Show flagged videos and detector output.

Not ground truth

The screenshot illustrates UI behavior, not validated detection accuracy.

What needs to be measured before deployment?

The detector is a hypothesis until benchmarked

Data

Create a provenance-backed, human/AI animation benchmark across styles and generators.

Model

Compare Hive, Skyra, DeMamba and others. Test held-out generators, precision/recall and latency.

Decision rule

Compare one-frame vs multi-frame threshold. Prioritize false-positive control and mixed-content cases.

Key systems trade-off: detection quality × inference cost × blocking latency.

Roadmap: make it a defensible Internet measurement

NOW

Browser prototype

3 sampled frames + Hive + user override

NEXT

Validated detector

Labeled animation benchmark + held-out-generator tests

THEN

**Exposure
measurement**

Measure profiles across YouTube / Shorts / Kids

Ethics: no real children's accounts, no commenting or ad clicks; separate video prevalence from impression exposure.

Takeaway

TinyShield gives families a potential user-controlled intervention and researchers a path to quantify exposure.

- Built: working extension with toggle, blocking and override.
- Missing: accuracy evidence for AI-generated animation.
- Next: validated detector and reproducible recommendation measurements.

Willing to work on TinyShield?

Contact: sonal.khosla@hpi.de

We are recruiting a PhD student

Contact: vaibhav.bajpai@hpi.de

References and caveats (backup)

1. NYT (26 Feb 2026), “How A.I.-Generated Videos Are Distorting Your Child’s YouTube Feed.”
2. YouTube Help, “Disclosing use of GenAI content” (accessed Aug 2026).
3. Regulation (EU) 2022/2065, Digital Services Act, Article 28.
4. Regulation (EU) 2024/1689, AI Act Article 50; European Commission Article 50 FAQ.
5. European Commission, EU KIDS Act, COM(2026) 681; proposed 17 Sep 2026.
6. Fløvig et al., TinyShield, ACM IMC 2026 Student Workshop two-pager.

Status note: prototype ≠ validated classifier; planned logging and prevalence study ≠ completed measurement.