FeatureFool: Zero-Query Fooling of Video Models via Feature Map

Supplementary Material

Table 7. Top-1 accuracy (%) of C3D and I3D after pre-training.

Dataset	C3D	I3D
UCF-101	83.54%	61.70%
HMDB-51	66.77%	47.92%
Kinetics-400	59.50%	71.80%

7. More Details on Victim Models

Pre-trained Video Classifiers. To obtain the source-classification models used in our study, we first performed standard supervised pre-training on the training partitions of UCF-101 [55] and HMDB-51 [30]. For the substantially larger Kinetics-400 [29] dataset, we adopted the official checkpoints released by the MXNet [10] project instead of retraining. Owing to disparate preprocessing conventions across benchmarks, the two backbones operate under different spatio-temporal resolutions: I3D [7] ingests 32-frame clips at 224×224 pixels when evaluated on Kinetics-400, whereas C3D [64] and every other dataset/backbone pair processes 16-frame stacks of size 112×112. The resulting recognition accuracies are summarised in Table 7.

8. Computing Maximum-Optical-Flow Frame with Farneback Algorithm

With Farneback Algorithm [18], given an input video $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times 3}$, we first convert it to grayscale and downsample to $0.5\times$ resolution to reduce computational cost. Let $\mathbf{X}_t^{\text{gray}} \in \mathbb{R}^{h\times w}$ denote the t-th grayscale frame, where $h = \lfloor H/2 \rfloor$ and $w = \lfloor W/2 \rfloor$.

Polynomial Expansion. For each pixel $\mathbf{p} = (x, y)$ we fit a quadratic polynomial inside a 5×5 neighbourhood $\mathcal{N}(\mathbf{p})$ by weighted least squares:

$$I(\mathbf{q}) \approx \mathbf{q}^\top \mathbf{A}(\mathbf{p})\, \mathbf{q} + \mathbf{b}(\mathbf{p})^\top \mathbf{q} + c(\mathbf{p}), \qquad \mathbf{q} \in \mathcal{N}(\mathbf{p}), \tag{22}$$

where $\mathbf{A}(\mathbf{p}) \in \mathbb{R}^{2\times 2}$ is symmetric matrix, $\mathbf{b}(\mathbf{p}) \in \mathbb{R}^2$ is linear, and $c(\mathbf{p})$ is constant. Weights are given by a 2-D Gaussian window $w_\sigma(\mathbf{q}) = \exp\!\big(-\|\mathbf{q}-\mathbf{p}\|^2/(2\sigma^2)\big)$; OpenCV [2] uses $\sigma = 1.2$ by default.

Two-Frame Displacement Constraint. Let the polynomial coefficients of two successive frames be $(\mathbf{A}_{t-1}, \mathbf{b}_{t-1})$ and $(\mathbf{A}_t, \mathbf{b}_t)$. Under the local translation assumption $\mathbf{A}_{t-1} \approx \mathbf{A}_t \triangleq \mathbf{A}$, the displacement vector $\mathbf{d}(\mathbf{p}) = [\Delta u, \Delta v]^\top$ satisfies

$$\mathbf{A}(\mathbf{p})\, \mathbf{d}(\mathbf{p}) = \frac{1}{2}\big[\mathbf{b}_{t-1}(\mathbf{p}) - \mathbf{b}_t(\mathbf{p})\big]. \tag{23}$$

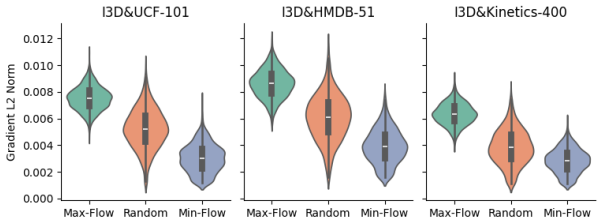

Figure 12. Normalised GB-gradient L_2-norm distributions across frames for three C3D-trained datasets. The distributions of Max-Flow frames are consistently shifted toward higher gradient magnitudes, validating their use as a proxy for the most model-sensitive locations in a black-box setting.

$$\mathbf{d}(\mathbf{p}) = \Big[\sum_{\mathbf{q}\in\Omega} w\mathbf{A}^\top\mathbf{A}\Big]^{-1} \sum_{\mathbf{q}\in\Omega} w\mathbf{A}^\top \frac{\mathbf{b}_{t-1} - \mathbf{b}_t}{2}. \tag{24}$$

Flow Magnitude. For the frame pair $(t-1, t)$ we obtain the dense flow field $\mathcal{F}_t(\mathbf{p}) = \mathbf{d}(\mathbf{p})$ and compute its average magnitude:

$$m_t = \frac{1}{hw}\sum_{\mathbf{p}\in\Omega} \|\mathcal{F}_t(\mathbf{p})\|_2 = \frac{1}{hw}\sum_{x,y} \sqrt{(\Delta u)^2 + (\Delta v)^2}. \tag{25}$$

Boundary handling: $m_0 = m_1$ and $m_T = m_{T-1}$.

Maximum-Optical-Flow Frame Index. Finally we select

$$t^* = \arg\max_{t=0,\ldots,T} m_t. \tag{26}$$

This frame is used by FEATUREFOOL for Guided Backpropagation.

9. More Experiments Results

9.1. Distributions of Flow frames.

As evidenced on the I3D [7]-family models in Figure 12, the gradient-norm distribution of Max-Flow frames is markedly shifted above those of Random- or Min-Flow frames. Therefore, we harvest a stronger decision-sensitive pattern from the I3D source network; broadcasting this pattern as a universal, motion-aligned perturbation across every frame of the victim video effectively misleads the black-box classifier without querying its gradients.

9.2. Variants Performance.

Table 8 and Table 9 report the results of variants on HMDB-51 [30] and Kinetics-400 [29]. Numerical results show that

Table 8. Attack performance comparison of FeatureFool variants on HMDB-51.

Model	Attack	HMDB-51			
		ASR↑	TI↓	SSIM↑	PSNR↑
C3D	FeatureFool-random	56%	3.4134	0.8619	29.1335
	FeatureFool-full	68%	3.7514	0.8637	**30.1103**
	FeatureFool	**72%**	**3.6821**	**0.8755**	28.9602
I3D	FeatureFool-random	62%	3.9355	**0.8955**	29.0109
	FeatureFool-full	67%	3.9521	0.8734	28.4061
	FeatureFool	**73%**	**4.3467**	0.8861	**29.4111**

Table 9. Attack performance comparison of FeatureFool variants on Kinetics-400.

Model	Attack	Kinetics-400			
		ASR↑	TI↓	SSIM↑	PSNR↑
C3D	FeatureFool-random	55%	3.9100	0.8650	**29.8505**
	FeatureFool-full	61%	**3.4200**	0.8710	28.9554
	FeatureFool	**70%**	3.7553	**0.8864**	28.5624
I3D	FeatureFool-random	59%	4.1200	**0.8840**	29.6508
	FeatureFool-full	66%	3.7800	0.8810	29.7026
	FeatureFool	**72%**	**3.6594**	0.8631	**30.2497**

SSIM [70] and PSNR are still influenced by the frame-selection strategy, yet ASR exhibits a consistent trend, confirming the effectiveness of maximum-optical-flow selection.

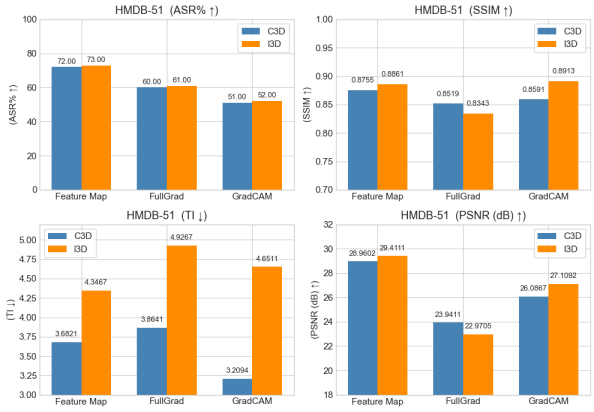

Figure 13. The performance of different noise types on HMDB-51.

9.3. Perturbation Selection.

Figures 13 and Figure 14 report the behavior of different noise variants on HMDB-51 and Kinetics-400. Again, the feature-map noise achieves the best trade-off between ASR and video quality: it delivers the highest ASR in all cases, benefiting from the finer-grained information it carries compared with other attention maps, and maintains superior adversarial-sample quality in most scenarios.

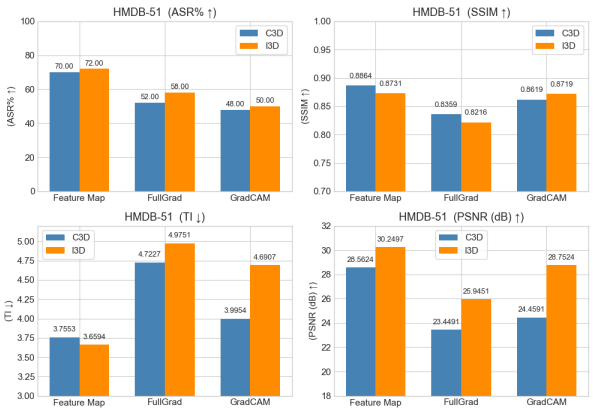

Figure 14. The performance of different noise types on Kinetics-400.

9.4. Cross-Evaluation across Noise Types.

We also conducted statistics on the cross-model-architecture and cross-dataset performance of FullGrad [57] and GradCam [51], and averaged the cross-performance. As shown in Figure 16, perturbations generated with the feature map achieve superior attack performance, especially on all ASR-related metrics. This benefit largely stems from the fact that feature maps carry finer-grained, semantically meaningful representations than the other two maps, enabling stronger yet equally imperceptible perturbations.

9.5. Trade-off between Stealthiness and ASR.

We explore the performance of FEATUREFOOL at $\alpha \in \{0.1, 0.4, 0.8, 1.0\}$. Tables 10, Tables 11 and Tables 12 present additional results across multiple datasets and models under these settings. Consistently across all splits, increasing α intensifies the feature-map injection, thereby exerting a stronger influence on the clean video and misleading the classifier. This gain in attack strength, however, comes at the cost of visual quality; higher α values visibly degrade SSIM and PSNR. Consequently, one must trade off stealthiness against ASR when setting the attack magnitude. Figures 15 illustrates the impact of α on UCF-101, HMDB-51, and Kinetics-400 with the corresponding C3D and I3D models. Figure 17 shows the visual comparison.

9.6. Hallucination Showcase.

We further observe that FEATUREFOOL can trigger hallucinations in Video-LLMs. An example is given in Figure 18: the injection of external features causes the model to output numerous irrelevant sentences, highlighted in red. One possible explanation is that the stealthy yet powerful feature perturbations introduced by FEATUREFOOL shift the clean video's feature space, forcing the Video-LLM to interpret the video within an adversarial space and ultimately producing hallucinations.

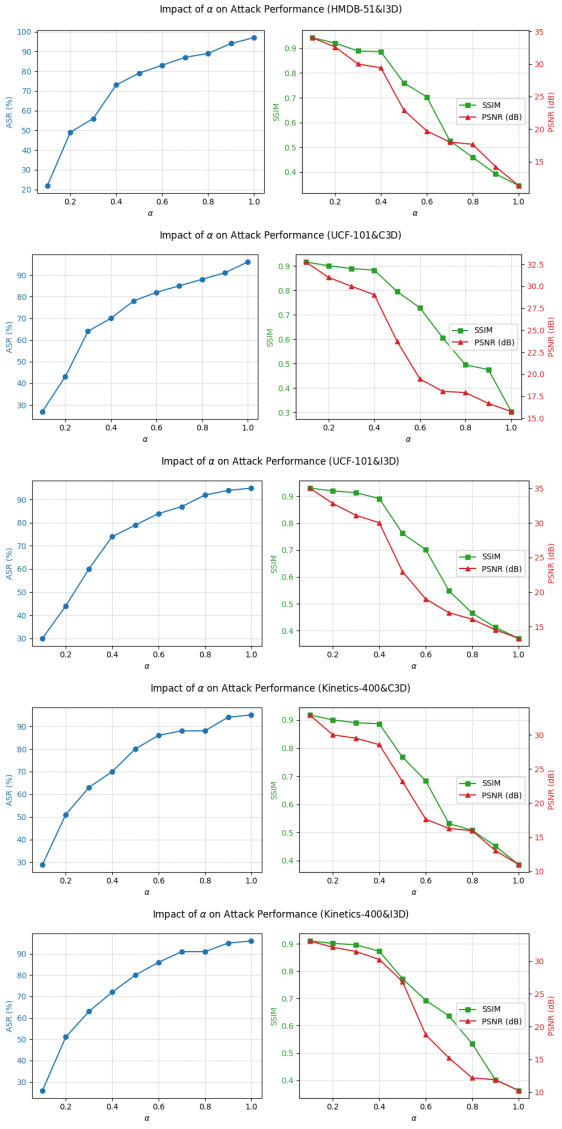

Figure 15. More results about impact of different α intensities on ASR, SSIM and PSNR.

Table 10. Attack performance with different α on UCF-101

Model	Attack	UCF-101			
		ASR↑	TI↓	SSIM↑	PSNR↑
C3D	FeatureFool (α=0.1)	27%	2.4244	0.9169	32.7569
	FeatureFool (α=0.4)	70%	3.1664	0.8834	29.0297
	FeatureFool (α=0.8)	88%	6.7056	0.4953	17.9106
	FeatureFool (α=1.0)	96%	10.0351	0.3044	15.7651
I3D	FeatureFool (α=0.1)	30%	2.1954	0.9303	35.0024
	FeatureFool (α=0.4)	74%	3.2937	0.8914	30.0137
	FeatureFool (α=0.8)	91%	5.9961	0.4657	16.1037
	FeatureFool (α=1.0)	95%	8.9421	0.3720	13.3473

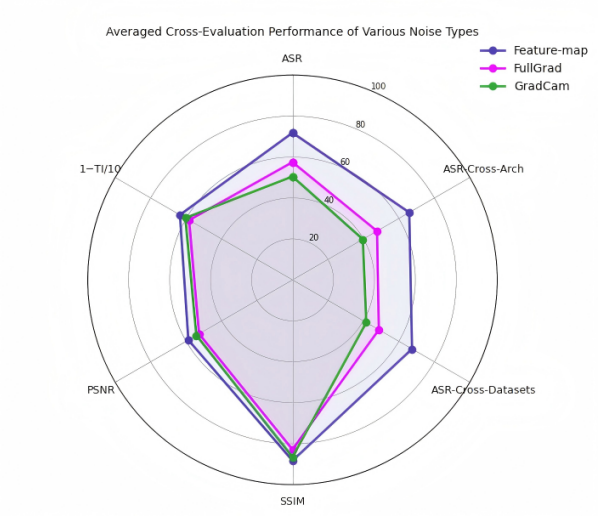

Figure 16. Radar-chart comparison of averaged performance across noise types. ASR: vanilla attack-success rate; ASR-Cross-Arch: cross-model transfer; ASR-Cross-Datasets: cross-dataset transfer; TI inverted for consistency. Feature-map perturbations (purple) consistently enclose the other polygons, demonstrating superior performance.

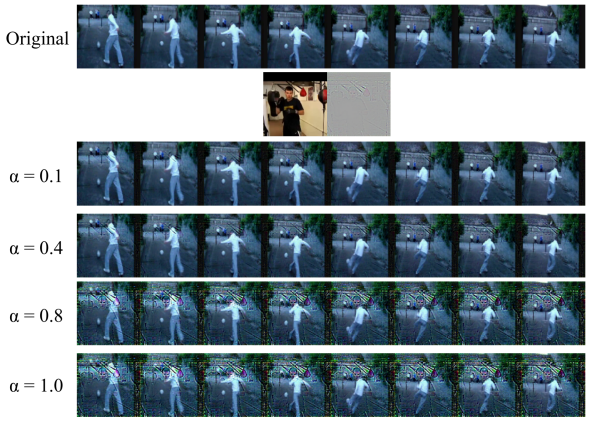

Figure 17. Impact of different α intensities on video appearance.

Table 11. Attack performance with different α on HMDB-51

Model	Attack	HMDB-51			
		ASR↑	TI↓	SSIM↑	PSNR↑
C3D	FeatureFool (α=0.1)	19%	3.0244	0.9467	33.9161
	FeatureFool (α=0.4)	72%	3.6821	0.8755	28.9602
	FeatureFool (α=0.8)	90%	6.4592	0.4682	15.1267
	FeatureFool (α=1.0)	96%	9.6651	0.3756	14.1134
I3D	FeatureFool (α=0.1)	22%	2.9181	0.9431	34.0624
	FeatureFool (α=0.4)	73%	4.3467	0.8861	29.4111
	FeatureFool (α=0.8)	89%	6.6304	0.4592	17.6642
	FeatureFool (α=1.0)	97%	8.2304	0.3450	11.3207

10. Details about Metrics.

Temporal Inconsistency (TI). Ruder et al. [50] minimise the squared warping residual

$$\mathcal{L}_{\text{temp}} = \sum c_k (x_k - \omega_k)^2 \tag{27}$$

Table 12. Attack performance with different α on Kinetics-400

Model	Attack	Kinetics-400			
		ASR↑	TI↓	SSIM↑	PSNR↑
C3D	FeatureFool (α=0.1)	29%	2.0244	0.9177	32.9119
	FeatureFool (α=0.4)	70%	3.7553	0.8864	28.5624
	FeatureFool (α=0.8)	88%	6.0482	0.5079	16.1109
	FeatureFool (α=1.0)	95%	10.6271	0.3864	11.0034
I3D	FeatureFool (α=0.1)	26%	3.2021	0.9105	33.0755
	FeatureFool (α=0.4)	72%	3.6594	0.8731	30.2497
	FeatureFool (α=0.8)	91%	6.4201	0.5346	12.1782
	FeatureFool (α=1.0)	96%	7.6891	0.3628	10.2679

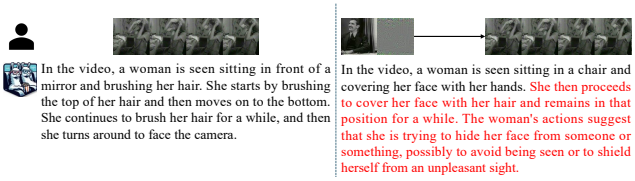

Figure 18. Hallucinations induced after FEATUREFOOL perturbation.

to stabilise stylised videos. We reuse their pipeline but measure the residual on adversarial videos. First, define the occlusion-weighted ℓ_1 error between any two frames

$$\mathcal{E}(x_t, x_m) = \frac{1}{HWC} \sum_{c=1}^{C} O_{t,m}^{(c)} \odot \left| x_t^{(c)} - \mathcal{W}(x_m^{(c)}) \right|, \quad (28)$$

where $\mathcal{W}$ is the DeepFlow [73] backward warp and $O_{t,m}$ the forward–backward consistency mask. Averaging over the whole clip yields the Temporal-Inconsistency index

$$\mathrm{TI} = \frac{1}{2(T-1)} \sum_{t=2}^{T} \Big[\mathcal{E}(x_t, x_1) + \mathcal{E}(x_t, x_{t-1}) \Big]. \quad (29)$$

Lower TI $\Rightarrow$ smoother motion; higher TI $\Rightarrow$ adversarial flicker.

Structural Similarity (SSIM). SSIM [70] assesses perceptual fidelity by comparing local luminance, contrast and structure between the clean video $\mathbf{X}$ and the adversarial video $\mathbf{X}_{\mathrm{adv}}$ frame-wise, then averaging over time:

$$\mathrm{SSIM}(\mathbf{X}, \mathbf{X}_{\mathrm{adv}}) = \frac{1}{T} \sum_{t=1}^{T} \mathrm{SSIM}(\mathbf{X}_t, \mathbf{X}_{\mathrm{adv},t}) \in [-1, 1], \quad (30)$$

where 1 indicates perfect visual match.

Peak Signal-to-Noise Ratio (PSNR). PSNR is computed on the ℓ_2 error of the 8-bit pixel space:

$$\mathrm{PSNR}(\mathbf{X}, \mathbf{X}_{\mathrm{adv}}) = 10 \log_{10} \frac{255^2}{\mathrm{MSE}(\mathbf{X}, \mathbf{X}_{\mathrm{adv}})} \ [\mathrm{dB}], \quad (31)$$

with $\mathrm{MSE} = \frac{1}{CHWT} \|\mathbf{X} - \mathbf{X}_{\mathrm{adv}}\|_2^2$. Higher PSNR (lower MSE) implies smaller perturbation energy.

11. Introduction to Video Datasets

UCF-101 [55] comprises 13,320 realistic videos distributed across 101 sport and daily-life categories. The collection is recorded at 25 fps with a spatial resolution of 320 × 240; most clips are 5–10 s long and depict nearly static scenes with stable camera motion.

HMDB-51 [30] provides 6,849 video clips from 51 action classes extracted YouTube, google and public databases. The dataset emphasises natural human motions (e.g., walk, wave, smile) under severe illumination changes, camera jitter and partial occlusions.

Kinetics-400 [29] is a large-scale corpus that contains $\approx$ 240 k training videos and 20 k validation videos spanning 400 human actions. Clips are sourced from YouTube at 25 fps with an average duration of 10 s; the action taxonomy covers fine-grained motions such as "playing violin" or "mopping floor".

Real-Life Violence Situations Dataset [17] contains 2,000 YouTube clips, half capturing diverse street-fight scenes and half everyday non-violent actions, that serve as realistic positive and negative samples for violence detection.

UCF-Crime Dataset [59] is the first large-scale dataset for real-world anomaly detection, offering 128 hours of untrimmed surveillance video that covers 13 realistic anomalies such as abuse, fighting, robbery and vandalism.

12. Important symbols

Table 13 lists the symbols frequently used in the main paper for quick reference.

Table 13. Frequently-used symbols in the main paper.

Symbol	Description
$\mathbf{X} \in \mathbb{R}^{T\times C\times H\times W}$	clean video clip
$\mathbf{X}^{\mathrm{att}}$	video that produces feature map (perturbation)
$\mathbf{X}_{\mathrm{adv}}$	adversarial video
δ	universal perturbation ($\|\delta\|_\infty \le \varepsilon$)
t^*	index of the Max-Optical-Flow frame
$\mathbf{G} \in \mathbb{R}^{H\times W\times C}$	Guided-Backprop feature map
α	injection strength of $\mathbf{G}$
$\phi(\cdot;\theta)$	3D-CNN classifier (C3D / I3D)
ASR	attack success rate (%)
TI	temporal-inconsistency
(S)	the source model in cross-evaluation
(V)	the victim model in cross-evaluation

13. Ethics Statement

Experiments on violence [17], crime [59] and pornography clips are conducted solely to evaluate model safety.

Adversarial perturbations do not create or intensify harmful content; Source videos are public, de-identified, audio-removed, and never re-distributed in adversarial form. Only aggregate metrics are reported. We encourage follow-up work on countermeasures and explicitly discourage any malicious reuse.

14. Future Work

FEATUREFOOL demonstrates effective and strong performance only zero-query in untargeted attacks within the video domain. Looking forward, future attention will shift to leveraging feature-map priors for targeted video attacks under zero- or few-query budgets. Moreover, we also care about the hallucinations (Sec. 9.6) in Video-LLMs that caused by feature-map injection, and investigate whether the hallucinations arise from attention sink [69] induced by the feature maps produced by FEATUREFOOL. Most importantly, future work will delve into more effective defenses against this class of feature-based, stealthy perturbations.