Frame Selection Strategies for Video Deepfake Detection: Benchmarking Accuracy and Runtime Trade-Offs

Artūras Serackis *, Mindaugas Jankauskas, Anastasija Grubinskienė, Vytautas Abromavičius

Posted Date: 21 April 2026

doi: 10.20944/preprints202604.1520.v1

Keywords: deepfake detection; frame selection; landmark-based sampling; reusable frame cache; frame-based detectors

 doi:10.20944/preprints202604.1520.v1

Article

Frame Selection Strategies for Video Deepfake Detection: Benchmarking Accuracy and Runtime Trade-Offs

Artūras Serackis *, Mindaugas Jankauskas, Anastasija Grubinskienė and Vytautas Abromavičius

Vilnius Gediminas Technical University, Lithuania
* Correspondence: arturas.serackis@vilniustech.lt; Tel.: +37-05-274-4765

Abstract

Deepfake detection from images and videos has evolved from artifact-specific convolutional baselines toward more generalizable, cross-dataset, and foundation-model-based approaches. The current work focuses on the efficiency and informativeness of frame selection itself, while keeping the downstream detectors fixed. The study compares twelve frame-selection heuristics ranging from simple baselines to landmark-aware strategies. Four pre-trained detectors were included in the present quantitative comparison: Self-Blended Images (SBI), Frequency-Enhanced Self-Lendered Images (FSBI), Generative Convolutional Vision Transformer (GenConViT), and GenD. The results show that GenD achieved the strongest average detector-level performance, with a mean frame-mean AUC of 0.9464. The best single validated configuration is GenD, yielding an AUC value of 0.9607 and a balanced accuracy of 0.9133. FSBI and SBI reached mean AUC values of 0.8953 and 0.8935, respectively, while GenD was the best general candidate. For SBI, the best validation configuration is Landmark cluster with 32 selected frames. GenD achieves the best AUC at the level of selection strategy. The present work demonstrates that inference-time frame selection is an important component of video-only deepfakes under constrained inference budgets.

Keywords: deepfake detection; frame selection; landmark-based sampling; reusable frame cache; frame-based detectors

1. Introduction

Frame-based deepfake detectors are commonly evaluated using a fixed number of frames per video, although the informativeness of individual frames is not uniform [1–3]. In practical video data, many frames are redundant, visually weak, or show only limited facial variation. As a result, frame selection should not be treated only as a preprocessing step, but rather as an experimental factor that can influence both classification performance and computational cost.

This issue is especially relevant for video-only deepfake detection pipelines that use pre-trained image-based or frame-aggregated models during inference [1–3]. Such models often rely on a relatively small subset of face frames extracted from the full clip. However, the choice of which frames are retained may affect the final video-level decision as much as the number of processed frames itself. In this context, reducing the frame budget can be valuable not only for faster inference and lower resource consumption, but also for clarifying whether the detector benefits more from temporal coverage, frame quality, motion, or geometric diversity of facial configurations.

The present study investigates this question using four pre-trained deepfake detectors: Self-Blended Images (SBI), Frequency-Enhanced Self-Blended Images (FSBI), Generative Convolutional Vision Transformer (GenConViT), and Deepfake Detection that Generalizes Across Benchmarks (GenD) [1–3]. The models have not been retrained. Instead, the study isolates the effect of inference-time of

 doi:10.20944/preprints202604.1520.v1

frame selection by evaluating how detector performance changes when only 2, 4, 8, 16 or 32 frames are selected from each video.

To address this problem, the study compares classical frame-selection heuristics with landmark-aware strategies. Classical baselines include uniform temporal sampling, visual diversity, quality-based ranking, motion-based sampling, shot-aware selection, face-utility scoring, TP-guided selection, and random sampling. The landmark-aware family uses pose-compensated facial landmark geometry to diversify the selected frames according to facial configuration rather than only visual appearance. In addition, hybrid variants combine landmark-based diversification with quality or motion cues.

The second contribution of the work is methodological. All detector evaluations are performed on a shared selection cache that materializes selected frames, face crops, and selection metadata before detector inference. Consequently, each detector sees the same selected frames for the same video, strategy, and frame budget. This removes repeated face extraction from the comparison loop, improves reproducibility, and makes detector-level differences easier to interpret. In this study, the cache is visual-only, which means that audio is not used in the benchmark. The emphasis on consistent inputs and controlled detector comparison aligns with the broader calls for standardized and reproducible deepfake evaluation pipelines [4].

Unlike a purely conceptual comparison, the present manuscript is aligned with the currently available benchmark outputs. The reported evidence corresponds to a 300-video evaluation set, 12 selection strategies, frame budgets of 2, 4, 8, 16, and 32, and a shared prepared cache comprising 18,000 video, strategy, and budget selections. At the current stage, complete quantitative results are available for four validated detectors, whereas the remaining configured baselines still await external checkpoints or successful reruns. For this reason, the manuscript emphasizes both the measured conclusions and the present experimental scope.

2. State of the Art

Deepfake detection from images and videos has evolved from artifact-specific convolutional baselines toward more generalizable, cross-dataset, and foundation-model-based approaches. Early studies established both the threat model and the first practical detection baselines. Initial works showed that deepfakes are challenging for classical biometric systems and standard forensic cues [5], while compact CNN detectors such as MesoNet [6], capsule-based models [7], and artifact-driven detectors based on face warping inconsistencies [8] demonstrated that manipulated faces can often be recognized from spatial artifacts. A major milestone was FaceForensics++, which standardized evaluation and accelerated the adoption of deep learning-based face forgery detection in common protocols [9].

A large line of work then focused on frequency-domain and spatial–frequency cues. F3Net showed that DCT-based frequency-aware representations are effective for face forgery detection, especially under compression [10]. SPSL emphasized phase information as a transferable cue for detecting up-sampling artifacts [11], while high-frequency feature learning further improved cross-dataset robustness [12]. More recent models combined spatial and frequency reasoning more explicitly, for example through graph-based relation learning across content and spectrum domains [13]. These studies established frequency analysis as one of the main alternatives to purely RGB-based detection.

Another important stream addressed local inconsistencies and fine-grained facial relations. Multi-attentional Deepfake Detection formulated the problem as fine-grained classification and used multiple attention heads to emphasize subtle local artifacts [14]. Pair-wise self-consistency learning (PCL-I2G) modeled source-feature inconsistency within forged faces [15]. Local Relation Learning introduced patch-level similarity patterns and RGB–frequency fusion to capture generalized local traces [16]. Learning second-order local anomalies further improved general face forgery detection by modeling higher-order inconsistencies [17]. Transformer-based detectors also entered this line of work, including UIA-ViT, which learns patch inconsistency without pixel-level forgery masks [18], and IID, which explicitly models implicit identity cues in face swapping [19].

 doi:10.20944/preprints202604.1520.v1

In parallel, several methods improved generalization through reconstruction, pseudo-fake generation, or self-supervised learning. Face X-rays exploited the boundaries of blending and remain one of the most influential image-based approaches for general detection of face forgery [20]. SBI replaced manipulation-specific synthetic data with self-blended images created from pristine faces and showed strong cross-dataset behavior [1]. RECCE combined reconstruction and classification in an end-to-end framework to obtain common features of genuine faces [21]. SLADD used self-supervised adversarial examples to improve generalization to unseen forgeries [22]. More recent training-oriented work continued this direction by revisiting whether deepfake data are strictly necessary during training [23] and by enriching pseudo-fake generation in the frequency domain through FreqBlender [24].

Recent state-of-the-art research has increasingly emphasized cross-dataset generalization and video-level robustness. AltFreezing addressed the imbalance between spatial and temporal evidence in video face forgery detection [25]. UCF proposed uncovering common forgery features that transfer better across benchmarks [26], while SeeABLE framed detection as a bounded one-class out-of-distribution problem using soft discrepancies [27]. Identity leakage was later identified as a major obstacle to generalization in binary deepfake classifiers [28]. LSDA expanded the effective forgery space by latent-space augmentation [29], and StyleGRU-style latent-flow modeling showed that temporal abnormalities in style latents are useful for video-level generalization [30]. At the same time, CLIP- and transformer-based detectors became increasingly prominent, including Forensics Adapter [31], GenConViT [2], and GenD, which demonstrated strong cross-benchmark performance with a parameter-efficient CLIP-based design [3].

Benchmarking and evaluation methodologies have also become a central part of the field. DeepfakeBench unified datasets, preprocessing, model implementations, and evaluation protocols for a large set of detectors [4]. DF40 also highlighted the need for more realistic and next-generation evaluation settings [32]. These benchmark efforts made it easier to compare detectors fairly, but they also revealed that most progress has been driven by detector architecture, training objectives, and cross-dataset generalization strategies rather than by systematic analysis of inference-time frame selection.

This observation is particularly relevant for the present study. Many successful deepfake detectors are image-based or frame-based models that ultimately operate on a fixed number of selected facial crops and then aggregate frame-level predictions into a video-level score [1–3,9,25]. However, the literature has devoted much more attention to designing stronger detectors than to studying which frames should be selected under a strict inference budget. In this sense, the current work complements the existing state of the art by shifting the focus from detector design to the efficiency and informativeness of frame selection itself, while keeping the downstream detectors fixed.

3. Materials and Methods

3.1. Study Configuration

The study considers binary video-level classification, where each input video is assigned to either the real or deepfake class using visual information only. The current benchmark uses 300 videos, 12 frame-selection strategies, and five frame budgets: 2, 4, 8, 16, and 32 selected frames per video. A total of 11 detectors were configured in the evaluation framework, but four detectors produced complete validated results in the present rerun: SBI, FSBI, GenConViT, and GenD.

Let v denote a video and let $\mathcal{C}(v) = \{x_1, x_2, \ldots, x_N\}$ denote its candidate face frames after preprocessing. For a frame-selection strategy π and frame budget $B \in \{2, 4, 8, 16, 32\}$, a subset

$$\mathcal{S}_{\pi,B}(v) \subseteq \mathcal{C}(v), \qquad |\mathcal{S}_{\pi,B}(v)| = B, \tag{1}$$

is selected for downstream inference. The same subset is then reused across all detectors evaluated under that video, strategy, and budget condition.

 doi:10.20944/preprints202604.1520.v1

3.2. Datasets

The experiments were conducted using three public deepfake benchmarks: FaceForensics++, Celeb-DF-v2, and Celeb-DF++. These datasets were selected because they represent different levels of manipulation realism, dataset scale, and evaluation difficulty.

FaceForensics++ is one of the most widely used benchmarks in face forgery detection and includes pristine and manipulated videos generated by several face manipulation methods [9]. It provides a standard experimental setting and remains an important reference dataset in the field.

Celeb-DF-v2 was introduced to provide more realistic and visually convincing deepfake videos with fewer obvious visual artifacts than previous datasets [33]. As a result, it is commonly used to evaluate whether detectors remain effective in more challenging and realistic conditions.

Celeb-DF++ is a more recent large-scale benchmark designed to support the evaluation of generalizable deepfake detection methods [34]. It extends the benchmark landscape to more diverse and difficult testing scenarios, making it especially relevant for studies focused on robustness and transferability.

Together, these datasets provide a suitable basis for analyzing whether reduced frame budgets and different frame-selection strategies preserve sufficient discriminative information across deepfake benchmarks of varying difficulty.

3.3. Compared Detectors

Four pre-trained deepfake detectors were included in the present quantitative comparison: SBI, FSBI, GenConViT, and GenD. The purpose of the study is not to retrain these models, but to evaluate how strongly their video-level behavior depends on frame budget and frame-selection policy. All detector weights were kept fixed during the experiments. Thus, the observed differences can be attributed to inference-time frame selection rather than to changes in optimization or training data.

3.4. Shared Selection Cache

A dedicated preparation stage materializes selected frames, selection manifests, and face crops before the detector benchmark. This design enables all detectors to operate on the same prepared visual inputs and avoids repeated frame extraction inside the evaluation loop. The current cache is visual-only, so audio streams are not used even if the source video originally contains them.

The prepared cache contains 18,000 video–strategy–budget selection rows in total. The final benchmark manifest contributes 100 videos from each source benchmark (Celeb-DF-v2, Celeb-DF++, and FaceForensics++), with each source balanced to 50 real and 50 fake clips. Across the five frame budgets, the benchmark therefore materializes 223,200 selected-frame positions before detector inference. These counts reflect the controlled benchmark inputs used in the present experiments.

3.5. Frame Selection Strategies

The study compares twelve frame-selection strategies that range from simple baselines to landmark-aware methods. Their role is to choose a subset of frames that is maximally informative under a strict frame budget. Table 1 summarizes the compared strategies.

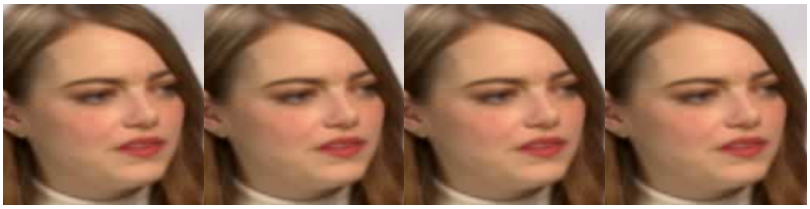
 doi:10.20944/preprints202604.1520.v1

Table 1. Frame-selection strategies used in the study.

Strategy	Description
Uniform	Samples frames at roughly equal temporal spacing and serves as the main simple baseline.
Diversity	Chooses frames that are visually most different from one another to avoid near-duplicates.
Quality	Ranks candidate face crops with a no-reference image-quality model and keeps the best-scoring ones.
Motion	Favors frames with stronger temporal change, where manipulations may be more visible.
Shot-aware	Spreads selected frames across shot-like segments to avoid collapsing onto a single local scene.
Face utility	Combines face size, centeredness, sharpness, and exposure into a task-oriented face usefulness score.
TP-guided	Uses supportive evidence from the prototype detector to prefer frames that look more helpful for true-positive detection.
Landmark diversity	After pose normalization, prefers frames whose landmark geometry differs the most.
Landmark cluster	Clusters pose-compensated landmark configurations and picks representatives from different geometric groups.
Landmark + quality	First diversifies by landmark clusters, then keeps the best-quality representative inside each cluster.
Landmark + motion	First diversifies by landmark geometry, then prioritizes representatives with stronger temporal change.
Random	Draws frames randomly and acts as a sanity-check baseline.

Besides the aggregate quantitative summaries reported later in the Results section, representative qualitative frame grids are retained here to show what the selection policies actually return on individual videos. These visual examples were generated from the same frame-selection pipeline and help illustrate how different strategies emphasize temporal coverage, image quality, and facial-geometry diversity on concrete clips. They are included as intuition-building examples rather than as separate evidence beyond the completed 300-video benchmark. The grids preserve an earlier 4/8/12/16 progression because those snapshots are visually easy to compare side by side, whereas the final quantitative benchmark reported in this manuscript uses the five budgets 2/4/8/16/32.

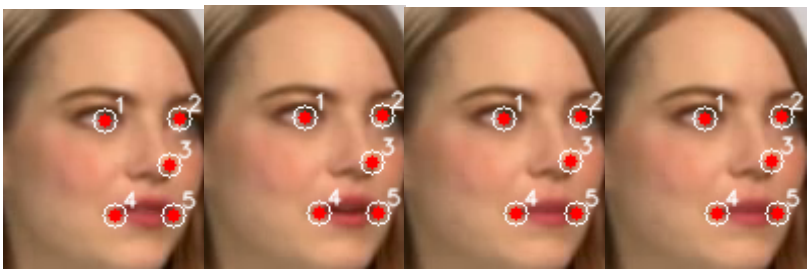
 doi:10.20944/preprints202604.1520.v1



(**a**) Uniform, 4f (**b**) Uniform, 8f (**c**) Uniform, 12f (**d**) Uniform, 16f

(**e**) Quality, 4f (**f**) Quality, 8f (**g**) Quality, 12f (**h**) Quality, 16f

(**i**) Landmark cluster, 4f (**j**) Landmark cluster, 8f (**k**) Landmark cluster, 12f (**l**) Landmark cluster, 16f



(**m**) Landmark + quality, 4f (**n**) Landmark + quality, 8f (**o**) Landmark + quality, 12f (**p**) Landmark + quality, 16f

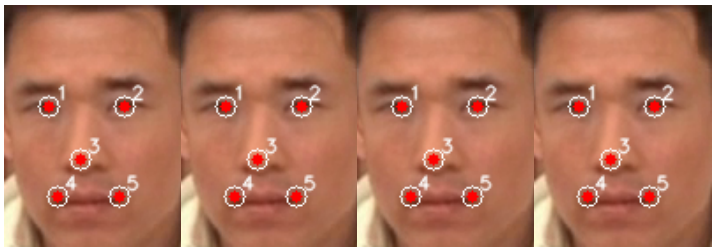
Figure 1. Illustrative frame-selection examples for a real video. Panels (**a**)–(**d**) show Uniform sampling for 4, 8, 12, and 16 selected frames. Panels (**e**)–(**h**) show Quality-based selection for the same budgets. Panels (**i**)–(**l**) show Landmark cluster selection. Panels (**m**)–(**p**) show Landmark + quality selection. The figure is retained as a qualitative visualization of how different policies emphasize different temporal locations and facial configurations on the same clip. The 12-frame column is illustrative only; the completed quantitative benchmark summarized later uses budgets of 2, 4, 8, 16, and 32.

 doi:10.20944/preprints202604.1520.v1

(**a**) Uniform, 4f (**b**) Uniform, 8f (**c**) Uniform, 12f (**d**) Uniform, 16f

(**e**) Quality, 4f (**f**) Quality, 8f (**g**) Quality, 12f (**h**) Quality, 16f

(**i**) Landmark cluster, 4f (**j**) Landmark cluster, 8f (**k**) Landmark cluster, 12f (**l**) Landmark cluster, 16f



(**m**) Landmark + quality, 4f (**n**) Landmark + quality, 8f (**o**) Landmark + quality, 12f (**p**) Landmark + quality, 16f

Figure 2. Illustrative frame-selection examples for a fake video. Panels (**a**)–(**d**) show Uniform sampling for 4, 8, 12, and 16 selected frames. Panels (**e**)–(**h**) show Quality-based selection for the same budgets. Panels (**i**)–(**l**) show Landmark cluster selection. Panels (**m**)–(**p**) show Landmark + quality selection. The figure is retained as a qualitative visualization of how different policies expose different facial appearances and geometric configurations in a manipulated clip. The 12-frame column is illustrative only; the completed quantitative benchmark summarized later uses budgets of 2, 4, 8, 16, and 32.

These strategies can be interpreted as using different notions of frame informativeness. Uniform and random sampling primarily control temporal coverage. Quality and face-utility methods prioritize technically usable face crops. Motion and shot-aware strategies emphasize temporal events and scene variation. The landmark-aware strategies instead focus on geometric diversity of facial configuration, which is especially relevant when a detector operates on individual face frames and may benefit from seeing a broader range of expressions, mouth shapes, or other facial deformations.

3.6. Video-Level Aggregation

For a detector $f_m(\cdot)$, each selected frame $x_i \in \mathcal{S}_{\pi,B}(v)$ yields a frame-level fake score

$$p_i = f_m(x_i). \tag{2}$$

 doi:10.20944/preprints202604.1520.v1

This formulation keeps the detector-agnostic aggregation rule and makes comparisons across frame budgets more direct.

3.7. Evaluation Protocol

The present quantitative analysis focuses primarily on the frame-mean AUC. Balanced accuracy is also reported for the best validated detector–strategy–budget configuration of each detector. The current manuscript should therefore be interpreted as a controlled comparison of frame-selection policies under a reduced-budget inference regime, rather than as a final ranking across all configured baselines, since four detectors currently have complete usable results.

4. Results

4.1. Overall Benchmark Summary

The current quantitative results are based on the completed 300-video benchmark and focus on frame-mean AUC as the main summary metric. Among the four validated detectors, GenD achieved the strongest average detector-level performance, with a mean frame-mean AUC of 0.9464. GenConViT followed with a mean AUC of 0.9247, while FSBI and SBI reached mean AUC values of 0.8953 and 0.8935, respectively. These results indicate that the detector ranking remains clearly separated even when all models are evaluated under the same shared frame-selection cache and reduced-budget inference protocol.

Table 2 summarizes the average frame-mean AUC obtained by each frame-selection strategy on the currently validated detectors.

Table 2. Average frame-mean AUC by frame-selection strategy on the currently validated detectors. Strategies are grouped into earlier heuristic baselines and landmark-aware variants.

Category	Strategy	Mean AUC
Heuristic baseline	Uniform	0.9109
Heuristic baseline	Diversity	0.9121
Heuristic baseline	Quality	0.9191
Heuristic baseline	Face utility	0.9141
Heuristic baseline	Random	0.9140
Heuristic baseline	Shot-aware	0.9190
Heuristic baseline	TP-guided	0.9147
Heuristic baseline	Motion	0.9084
Landmark-aware	Landmark cluster	0.9183
Landmark-aware	**Landmark + quality**	**0.9197**
Landmark-aware	Landmark diversity	0.9143
Landmark-aware	Landmark + motion	0.9156

At the frame budget level, the strongest average operating point across the validated detectors was 32 selected frames, with a mean frame-mean AUC of 0.9257. The remaining budgets followed a consistent ascending order: 16 frames reached 0.9218, 8 frames reached 0.9189, 4 frames reached 0.9129, and 2 frames reached 0.8957. Thus, the current benchmark does not show a reversal at higher frame budgets; however, the improvement from 8 to 32 frames is modest compared with the jump from 2 to 8 frames, which indicates diminishing returns rather than a strict need for dense frame usage.

At the level of the selection strategy, the best average performance was obtained by Landmark + quality, which reached a mean frame-mean AUC of 0.9197. The next strongest strategies were Quality (0.9191), Shot-aware (0.9190), and Landmark cluster (0.9183). The weakest average strategy in the current validated subset was Motion, with a mean AUC of 0.9084. The gap between the best landmark-aware and best heuristic strategies is therefore small but persistent, which suggests that geometric diversification is useful in the completed benchmark, while quality-aware heuristics remain nearly as strong.

 doi:10.20944/preprints202604.1520.v1

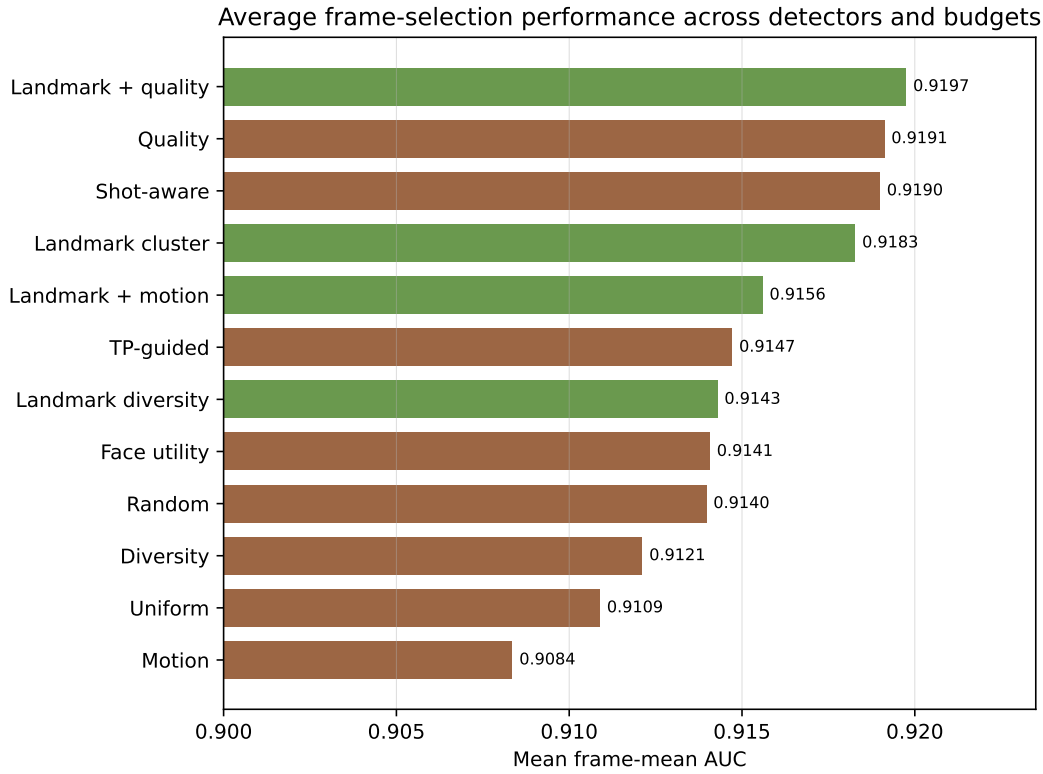


Figure 3. Average frame-mean AUC for each frame-selection strategy across all four validated detectors and all five frame budgets. Landmark-aware strategies are shown separately from earlier heuristic baselines in the generated PDF. The current benchmark places Landmark + quality at the top, with Quality and Shot-aware following closely behind.

4.2. Detector-Specific Best Configurations

Although landmark-aware methods were strongest on average across validated detectors, the best configuration still depended on the detector.

For GenConViT, the best validated configuration was Diversity with 32 selected frames, yielding an AUC of 0.9394 and a balanced accuracy of 0.7367. This result suggests that the model benefits from strong appearance diversity once the frame budget is allowed to increase in the higher-budget regime.

For GenD, the best validated configuration was Shot-aware with 32 selected frames, yielding an AUC of 0.9607 and a balanced accuracy of 0.9133. This is the strongest single frame-mean result among the currently validated detectors and indicates that the model benefits from a broader temporal spread once the budget is large enough to cover multiple local segments.

For FSBI, the best validated configuration was Uniform with 32 selected frames, producing an AUC of 0.9147 and a balanced accuracy of 0.8467. The frequency-enhanced variant therefore closed most of the gap between SBI and the stronger CLIP-style detectors, but still benefited most from the largest tested frame budget.

For SBI, the best validated configuration was the Landmark cluster with 32 selected frames, yielding an AUC of 0.9097 and a balanced accuracy of 0.8533. Compared with the other detectors, SBI remained the weakest average detector overall, yet it still crossed the 0.90 AUC mark when paired with a landmark-aware strategy at the largest tested budget.

 doi:10.20944/preprints202604.1520.v1

Table 3. Best validated configuration for each detector on the completed 300-video benchmark.

Detector	Best strategy	Frames	AUC	Balanced accuracy
GenD	Shot-aware	32	0.9607	0.9133
GenConViT	Diversity	32	0.9394	0.7367
FSBI	Uniform	32	0.9147	0.8467
SBI	Landmark cluster	32	0.9097	0.8533

4.3. Effect of Frame Budget

The detector-wise average AUC trends across frame budgets showed a consistent upward pattern. GenConViT reached mean AUC values of 0.9081, 0.9210, 0.9251, 0.9329, and 0.9365 for 2, 4, 8, 16, and 32 frames, respectively. GenD reached 0.9310, 0.9410, 0.9501, 0.9537, and 0.9564. FSBI reached 0.8686, 0.8936, 0.9042, 0.9027, and 0.9077, while SBI reached 0.8751, 0.8961, 0.8960, 0.8981, and 0.9024. Thus, all four validated detectors benefited from larger frame budgets, but the gains became progressively smaller beyond 8 or 16 frames. This finding is important from a computational perspective because it shows that a 32-frame budget is strongest in absolute terms, while smaller budgets still retain most of the achievable AUC.

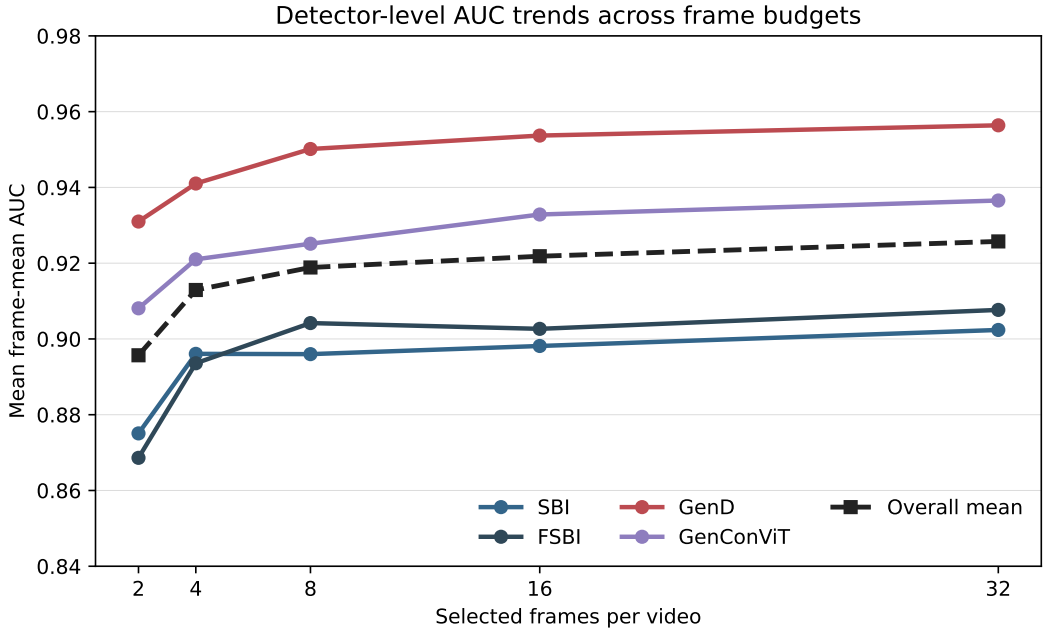


Figure 4. Detector-wise mean frame-mean AUC across frame budgets. The generated PDF includes the four validated detectors and the overall mean. GenD remains the strongest detector at every budget, while all detectors show diminishing but positive gains as the frame budget increases from 2 to 32.

To test whether the small gaps between the higher budgets reflect a stable tendency rather than noise in a single aggregate table, a bootstrap analysis was performed directly on the video-level prediction files of the 48 validated detector–strategy configurations available at each budget. The resulting 95% confidence intervals for the overall mean AUC were 0.9149–0.9237 at 8 frames, 0.9169–0.9263 at 16 frames, and 0.9209–0.9307 at 32 frames. These intervals overlap substantially, which supports the interpretation that the completed benchmark exhibits diminishing returns rather than a sharp transition between the higher budgets.

A complementary delta-to-32 analysis makes the same point from the detector perspective. Relative to the 32-frame baseline, the overall mean AUC loss is only 0.0069 at 8 frames and 0.0039 at 16 frames. The largest 8-to-32 drop is observed for GenConViT (0.0114), whereas FSBI loses only

 doi:10.20944/preprints202604.1520.v1

0.0035 on average. Thus, the additional gains above 8 or 16 frames are real but detector-dependent and comparatively small.

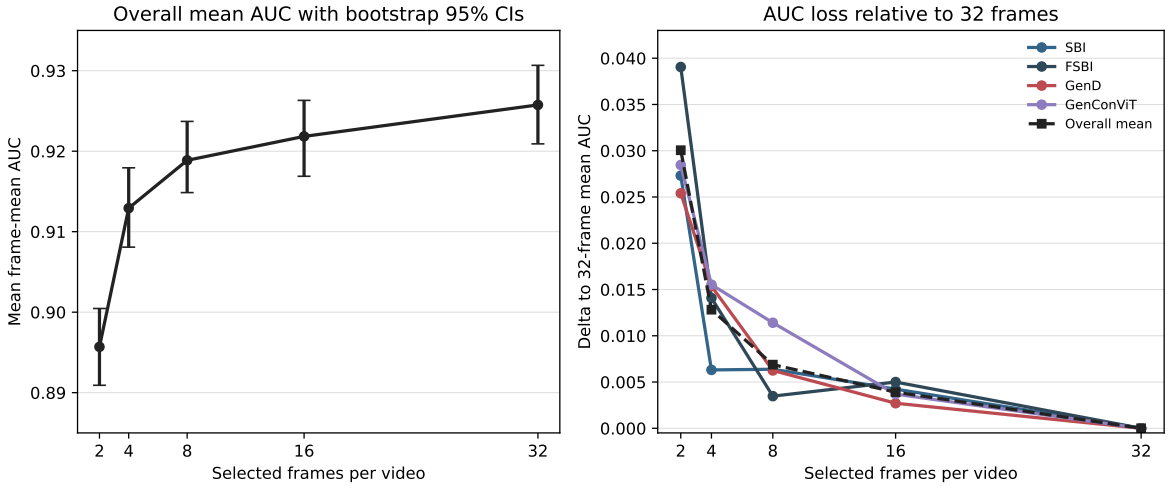


Figure 5. Frame-budget interpretation from two complementary views. Left: overall mean frame-mean AUC with bootstrap 95% confidence intervals, computed from video-level prediction files across the validated detector–strategy configurations. Right: detector-specific AUC loss relative to the corresponding 32-frame mean. Together, the panels show that 32 frames remain strongest in absolute terms while the 8- and 16-frame settings already recover most of the attainable accuracy.

4.4. Selection-Stage Runtime

The completed benchmark also preserved wall-clock information about the shared SBI-based selection stage. All wall-clock timings reported in this section were obtained on the same server equipped with an NVIDIA H100 NVL GPU (95.8 GB VRAM), an Intel Xeon Gold 6548Y+ CPU (16 cores), and 125 GB of system memory, running Ubuntu 24.04.2 LTS with PyTorch 2.5.1 and CUDA 12.4. Consequently, the absolute runtimes should be interpreted as hardware-dependent, whereas the relative comparisons remain directly comparable within this benchmark. The entire 300-video benchmark required 189,936.7 seconds of wall-clock time from start to summary generation, or approximately 52.8 hours. At a finer level, the per-strategy SBI runtime logs show that the selection stage scales sharply with the frame budget. Averaged across the 12 strategies, the total selection-stage runtime increased from 295.7 seconds at 2 frames to 587.8 seconds at 4 frames, 1195.0 seconds at 8 frames, 2303.1 seconds at 16 frames, and 4789.5 seconds at 32 frames.

This runtime view helps to interpret the practical meaning of the AUC trends. Relative to the 32-frame setting, the 8-frame budget already recovers about 99.3% of the mean frame-mean AUC while using only about 25% of the average recorded SBI selection-stage runtime. The 16-frame budget recovers about 99.6% of the mean frame-mean AUC while using about 48% of that runtime. Therefore, although 32 frames remain strongest in absolute terms, the smaller mid-range budgets remain attractive efficiency-oriented operating points.

Within the runtime matrix, Uniform remains the cheapest strategy at every budget, while the more complex hybrid and guided policies become progressively more expensive as the budget grows. At 32 frames, for example, Uniform required 1124.6 seconds, whereas Landmark + motion reached 7133.4 seconds and Landmark + quality reached 5531.4 seconds. The timing logs are dominated by prototype generation; the recorded evaluation overhead in these SBI runtime files remained negligible by comparison.

 doi:10.20944/preprints202604.1520.v1

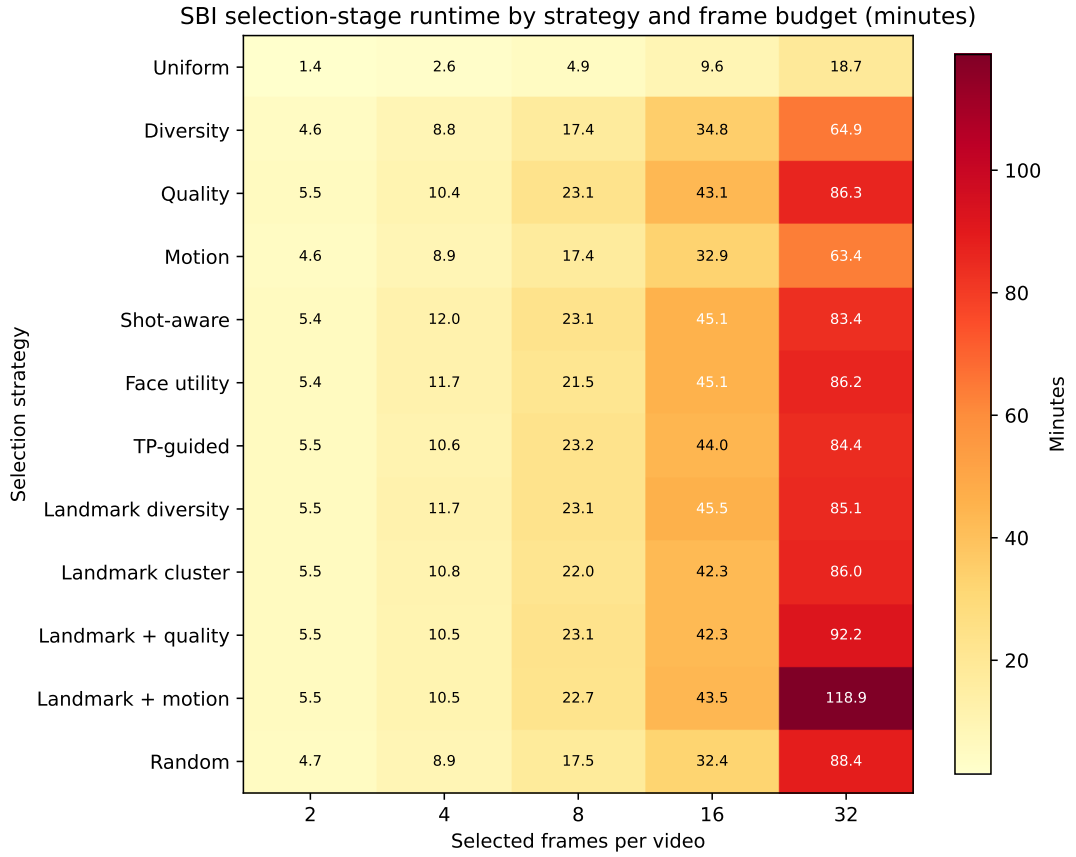


Figure 6. SBI selection-stage runtime by strategy and frame budget, measured on the completed 300-video benchmark. Each cell reports total runtime in minutes for the corresponding strategy–budget pair. The figure highlights the steep growth in selection-stage cost at higher budgets and shows that the more complex hybrid and guided strategies become substantially more expensive than Uniform as the frame budget increases.

4.5. Strategy Cost Benefit Trade-Off at 32 Frames

The heatmap is useful for understanding raw runtime growth, but the most decision-relevant view is the cost–benefit trade-off at the largest tested budget. At 32 frames, Uniform remained by far the cheapest evaluated SBI selection policy at 18.7 minutes while still reaching a frame-mean AUC of 0.9080. Landmark cluster achieved the strongest SBI result at the same budget, with an AUC of 0.9097, but required 86.0 minutes. Landmark + quality was nearly tied in accuracy at 0.9095 while taking 92.2 minutes. By contrast, the Landmark + motion reached only 0.8952 AUC despite requiring 118.9 minutes, making it a particularly expensive configuration without a matching accuracy return.

This comparison clarifies that the completed benchmark does not reward complexity uniformly. Some expensive selection strategies improve accuracy slightly, but the gains can be marginal relative to their wall-clock cost. In particular, the Pareto view shows that Uniform remains a strong efficiency baseline, Landmark cluster is a high-accuracy but much more expensive operating point, and Landmark + motion is dominated by cheaper alternatives with better or comparable performance. This type of trade-off analysis is important if frame selection is to be used in practical deployments rather than only as an offline benchmark variable.

 doi:10.20944/preprints202604.1520.v1

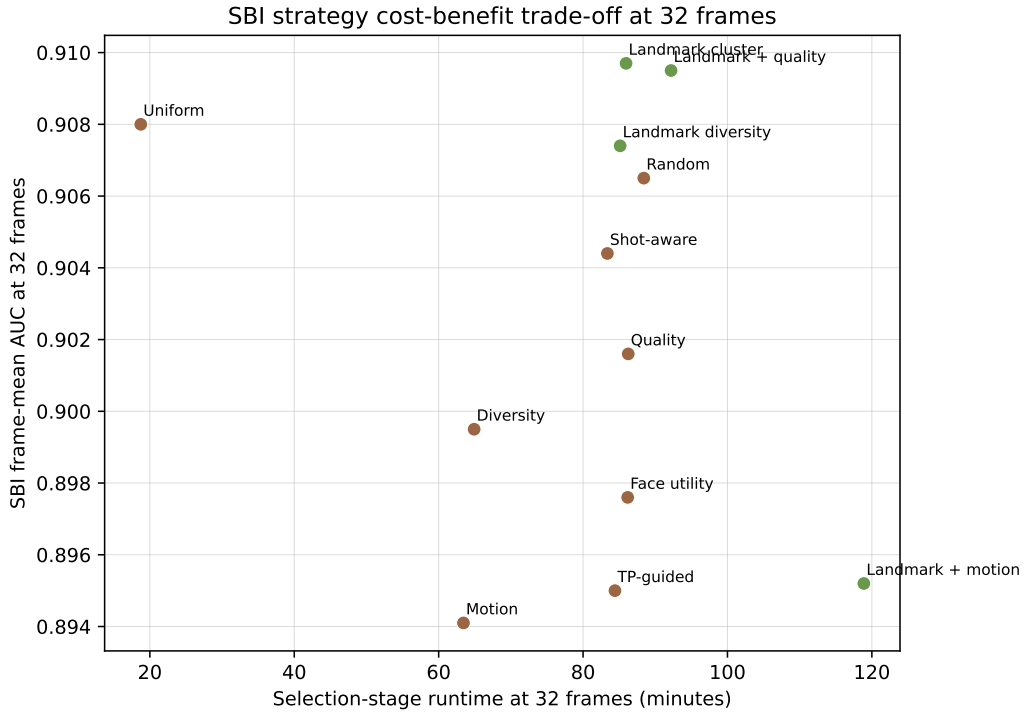


Figure 7. Cost–benefit trade-off of SBI frame-selection strategies at 32 selected frames. Each point combines the recorded selection-stage runtime and the corresponding frame-mean AUC. Uniform remains the cheapest competitive baseline, Landmark cluster provides the strongest SBI accuracy at a much higher cost, and Landmark + motion is expensive without a matching AUC gain.

The same observation can be summarized with a strict dominance criterion. At 32 frames, a strategy is considered dominated if another strategy achieves at least the same AUC at equal or lower runtime, with strict improvement in at least one of the two dimensions. Under this definition, ten of the twelve SBI strategies are dominated. Uniform dominates nine of them directly, while Landmark cluster dominates Landmark + quality by achieving slightly higher AUC at lower runtime. Consequently, only Uniform and Landmark cluster remain non-dominated in the 32-frame SBI selection-stage comparison.

Table 4. Strictly dominated SBI selection strategies at 32 selected frames.

Dominated strategy	AUC	Runtime (min)	Dominated by	AUC	Runtime (min)
Diversity	0.8995	64.9	Uniform	0.9080	18.7
Quality	0.9016	86.3	Uniform	0.9080	18.7
Motion	0.8941	63.4	Uniform	0.9080	18.7
Shot-aware	0.9044	83.4	Uniform	0.9080	18.7
Face utility	0.8976	86.2	Uniform	0.9080	18.7
TP-guided	0.8950	84.4	Uniform	0.9080	18.7
Landmark diversity	0.9074	85.1	Uniform	0.9080	18.7
Landmark + quality	0.9095	92.2	Landmark cluster	0.9097	86.0
Landmark + motion	0.8952	118.9	Uniform	0.9080	18.7
Random	0.9065	88.4	Uniform	0.9080	18.7

4.6. Detector Runtime Under Uniform Sampling

To complement the shared selection-stage timings, a separate post-benchmark wall-clock sweep was executed for the three external validated detectors under the Uniform strategy on the same

 doi:10.20944/preprints202604.1520.v1

300-video manifest. This gives a simple detector-level runtime baseline that is not confounded by the additional cost of more complex selection policies. Across all tested budgets, GenD was the slowest detector, while FSBI and GenConViT remained close to one another and exchanged the second place depending on the frame budget.

At 32 frames, the recorded wall-clock times were 19.8 minutes for FSBI, 24.6 minutes for GenD, and 21.7 minutes for GenConViT. At 16 frames, the corresponding times were 10.0, 12.9, and 11.8 minutes. Thus, the strongest detector in accuracy terms, GenD, is also the most computationally expensive of the three measured external detectors under this common baseline. This result is useful for deployment-oriented interpretation because it shows that the detector with the highest AUC is not automatically the most efficient operating choice.

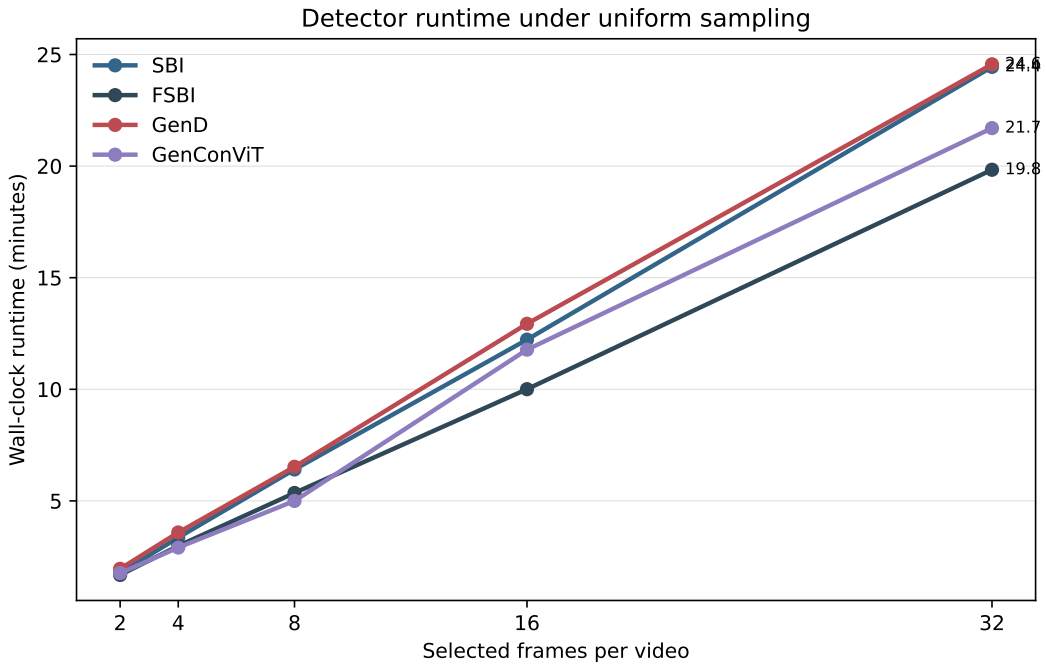


Figure 8. Wall-clock detector runtime under Uniform sampling across the five frame budgets on the completed 300-video benchmark. The figure reports the post-benchmark timing sweep for FSBI, GenD, and GenConViT and shows that GenD remains the slowest detector at every tested budget, while FSBI and GenConViT stay comparatively close.

4.7. Interpretation of Strategy Ranking

The average strategy ranking indicates that explicit geometric diversification is a strong principle for frame selection, but the completed benchmark also shows that landmark-aware and heuristic strategies now form a tight top tier rather than two clearly separated groups. Landmark + quality and Landmark cluster remain among the best general-purpose choices, yet Quality and Shot-aware are nearly tied with them. At the same time, the best per-detector strategy is not identical across models. Therefore, the present results support two complementary conclusions: first, landmark-aware strategies are strong general candidates for efficient deepfake detection; second, detector-specific interactions between frame-selection policy and model architecture still matter and should be analyzed separately.

 doi:10.20944/preprints202604.1520.v1

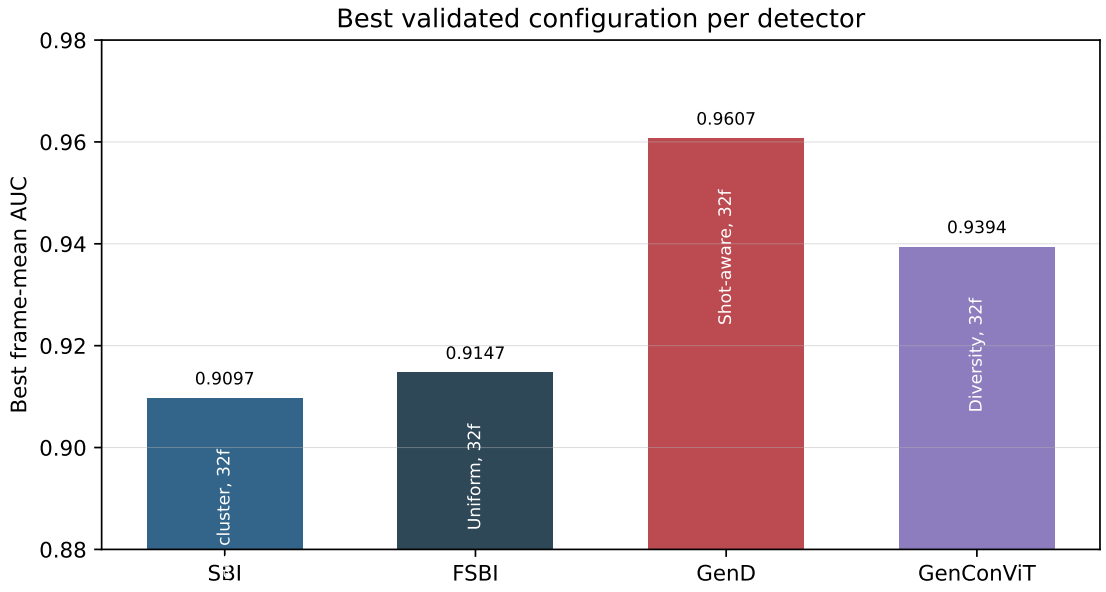


Figure 9. Best validated frame-mean AUC for each detector together with the corresponding strategy and frame budget. All four validated detectors achieve their strongest frame-mean AUC at 32 selected frames, but they do so with different selection policies.

5. Discussion

The completed benchmark revises the strongest version of the original low-budget hypothesis. In the present 300-video study, the best absolute average operating point is not 8 or 16 selected frames, but 32 selected frames. This is consistent across all four validated detectors. However, the more practically important observation is that the gains become small beyond the mid-range budgets. The average frame-mean AUC rises from 0.8957 at 2 frames to 0.9189 at 8 frames, then only to 0.9218 at 16 frames and 0.9257 at 32 frames. The bootstrap confidence intervals of the higher-budget means overlap substantially, which further supports a diminishing-returns interpretation rather than a strong qualitative separation between 8, 16, and 32 frames. Thus, 8 and especially 16 frames recover most of the attainable performance while requiring substantially fewer selected inputs than the largest tested budget.

This distinction matters for the interpretation of novelty. The present manuscript does not claim that fewer frames universally outperform larger budgets. Instead, its contribution is a controlled and detector-consistent quantification of the marginal returns of additional frames under a shared inference-time selection cache. By keeping the detector weights fixed and reusing the same selected inputs across SBI, FSBI, GenConViT, and GenD, the study isolates frame-selection effects from retraining effects. This allows the benchmark to answer a more specific question than most prior detector papers, namely how much performance is gained by moving from 2 to 4, 8, 16, and 32 selected frames under the same downstream conditions.

The strategy ranking also became more nuanced after the completed rerun. Landmark-aware selection remains strong, with Landmark + quality achieving the highest mean AUC and Landmark cluster remaining in the top tier. At the same time, the gap to the best heuristic strategies is small: Quality and Shot-aware are nearly tied with the best landmark-aware strategy in the validated subset. The runtime-aware Pareto view makes this even clearer: some sophisticated policies improve accuracy only marginally relative to Uniform, while others such as Landmark + motion become expensive without competitive AUC. Therefore, the evidence supports geometric diversification as a robust general principle, but it does not support an overly strong claim that landmark-aware methods dominate all alternatives under all conditions.

 doi:10.20944/preprints202604.1520.v1

Detector-specific interactions remain important. All four validated detectors achieve their best frame-mean AUC at 32 selected frames, yet the preferred strategy differs by model: Landmark cluster for SBI, Uniform for FSBI, Diversity for GenConViT, and Shot-aware for GenD. This indicates that frame selection should not be treated as a universal preprocessing choice detached from detector architecture. Instead, it should be tuned jointly with the detector when inference cost, latency, or deployment constraints matter.

These findings complement earlier deepfake detection studies, which mainly focus on detector design and usually adopt a fixed inference-time frame count [1–3]. In contrast, the present work shows that inference-time frame selection itself is a meaningful experimental variable that changes the cost–accuracy trade-off even when the detector is kept fixed. The shared selection cache is important in this context because it ensures that all detectors are compared on exactly the same selected inputs, and therefore makes detector-level differences easier to interpret.

The current conclusions should be interpreted within the scope of the benchmark. The quantitative evidence is based on 300 videos drawn from Celeb-DF-v2, Celeb-DF++, and FaceForensics++, and on four validated detectors, while several additional integrated baselines still await external checkpoints or compatibility fixes. The study is also visual-only and does not address audio or multimodal cues. Future work should therefore extend the comparison to a larger detector set, test the same strategies on broader datasets and compression conditions, and report explicit runtime and energy savings in addition to classification metrics.

The study indicates that frame selection is not merely a technical detail to reduce the input size. It is a meaningful design choice that shapes the final video-level decision, and its practical value lies less in beating the largest frame budget than in showing how close smaller budgets can come to the same performance under controlled conditions.

6. Conclusions

This study demonstrates that frame selection is an important component of deepfake detection with video-only under constrained inference budgets. In the completed 300-video benchmark with three source datasets and four validated detectors, the strongest average operating point is 32 selected frames. However, the benchmark also shows that most of the attainable performance is already recovered at 8 and 16 selected frames, and the bootstrap confidence intervals of the higher-budget means overlap substantially. Therefore, the practical question is not whether 32 frames are best in absolute terms but how much accuracy is sacrificed when substantially smaller budgets are used.

The benchmark also shows that detector ranking and frame-selection ranking should be interpreted separately. At the detector level, GenD is the strongest validated model, achieving a mean frame-mean AUC of 0.9464. At the strategy level, Landmark + quality is the strongest average selection method with a mean AUC of 0.9197, followed closely by Quality, Shot-aware, and Landmark cluster. However, the runtime-aware comparison shows that not every stronger or more complex strategy is cost-effective: Uniform remains a strong efficiency baseline, whereas some expensive hybrids do not yield proportionally better accuracy. Therefore, the study captures not only differences in detector quality but also the independent contribution of the selection policy itself under a controlled shared-cache protocol.

A further important conclusion is that landmark-aware strategies are strong general-purpose choices, but do not eliminate detector-specific behavior. The best detector-specific configuration is GenD with Shot-aware sampling at 32 frames, while the other detectors prefer different strategies, even though all peak at the same budget. This supports continued investigation of pose-compensated geometric diversification, but it also advocates detector-aware tuning rather than a one-size-fits-all frame-selection policy.

The present manuscript should be interpreted as a reproducible benchmark of inference-time frame selection rather than as a claim that low budgets universally outperform larger ones. Its main contribution is the controlled comparison itself: a shared prepared cache, 12 strategies, five frame

 doi:10.20944/preprints202604.1520.v1

budgets, and detector-consistent evaluation across Celeb-DF-v2, Celeb-DF++, and FaceForensics++. As additional checkpoints become available, the same benchmark can be extended directly without changing the prepared frame-selection data set, making the current study a useful baseline for future comparative evaluation.

Author Contributions: Conceptualization, A.S. and M.J.; methodology, A.S.; software, M.J.; validation, M.J. and A.G.; formal analysis, A.G.; investigation, M.J. and V.A.; resources, A.G.; data curation, A.G.; writing–original draft preparation, M.J.; writing–review and editing, A.S. and V.A.; visualization, M.J.; supervision, V.A.; project administration, A.G.; funding acquisition, A.S. All authors have read and agreed to the published version of the manuscript.

Funding: Research was conducted as part of the execution of Project "Mission-driven Implementation of Science and Innovation Programmes" (No. 02-002-P-0001) "Hybrid, Information, Psychological, Societal Threats handling system for public security domain practitioners, businesses, and education", funded by the Economic Revitalization and Resilience Enhancement Plan "New Generation Lithuania".

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area under the curve
CLIP	Contrastive Language-Image Pre-Training
CNN	Convolutional Neural Network
DCT	Discrete Cosine Transform
FSBI	Frequency-Enhanced Self-Blended Images
GenD	Deepfake Detection that Generalizes Across Benchmarks
GenConViT	Generative Convolutional Vision Transformer
IID	Implicit Identity Driven
PCL-I2G	Pair-wise self-consistency learning
RECCE	REConstruction-Classification lEarning
RGB	Red Green Blue
SBI	Self-Blended Images
SLADD	Self-supervised Learning of Adversarial Deppfake Detector
SPSL	Spatial-Phase Shallow Learning

References

1. Shiohara, K.; Yamasaki, T. Detecting deepfakes with self-blended images. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 18720–18729. https://doi.org/10.1109/CVPR52688.2022.01816.
2. Deressa, D.W.; Mareen, H.; Lambert, P.; Atnafu, S.; Akhtar, Z.; Van Wallendael, G. GenConViT: Deepfake video detection using generative convolutional vision transformer. *Applied Sciences* **2025**, *15*, 6622. https://doi.org/10.3390/app15126622.
3. Yermakov, A.; Cech, J.; Matas, J.; Fritz, M. Deepfake detection that generalizes across benchmarks. In Proceedings of the Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2026, pp. 773–783. https://doi.org/10.48550/arXiv.2508.06248.
4. Yan, Z.; Zhang, Y.; Yuan, X.; Lyu, S.; Wu, B. Deepfakebench: A comprehensive benchmark of deepfake detection. *arXiv preprint arXiv:2307.01426* **2023**. https://doi.org/10.48550/arXiv.2307.01426.
5. Korshunov, P.; Marcel, S. DeepFakes: A New Threat to Face Recognition? Assessment and Detection. *arXiv preprint arXiv:1812.08685* **2018**. https://doi.org/10.48550/arXiv.1812.08685.
6. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. MesoNet: a Compact Facial Video Forgery Detection Network. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7. https://doi.org/10.1109/WIFS.2018.8630761.

 doi:10.20944/preprints202604.1520.v1

7. Nguyen, H.H.; Yamagishi, J.; Echizen, I. Capsule-Forensics: Using Capsule Networks to Detect Forged Images and Videos. In Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2307–2311. https://doi.org/10.1109/ICASSP.2019.8682602.
8. Li, Y.; Lyu, S. Exposing DeepFake Videos By Detecting Face Warping Artifacts. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2019. https://doi.org/10.48550/arXiv.1811.00656.
9. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1–11. https://doi.org/10.1109/ICCV.2019.00009.
10. Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; Shao, J. Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues. In Proceedings of the Computer Vision – ECCV 2020, 2020, pp. 86–103. https://doi.org/10.1007/978-3-030-58610-2_6.
11. Liu, H.; Li, X.; Zhou, W.; Chen, Y.; He, Y.; Xue, H.; Zhang, W.; Yu, N. Spatial-Phase Shallow Learning: Rethinking Face Forgery Detection in Frequency Domain. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 772–781. https://doi.org/10.1109/CVPR46437.2021.00083.
12. Luo, Y.; Zhang, Y.; Yan, J.; Liu, W. Generalizing Face Forgery Detection With High-Frequency Features. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 16317–16326. https://doi.org/10.1109/CVPR46437.2021.01605.
13. Wang, Y.; Yu, K.; Chen, C.; Hu, X.; Peng, S. Dynamic Graph Learning With Content-Guided Spatial-Frequency Relation Reasoning for Deepfake Detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7278–7287. https://doi.org/10.1109/CVPR52729.2023.00703.
14. Zhao, H.; Wei, T.; Zhou, W.; Zhang, W.; Chen, D.; Yu, N. Multi-attentional Deepfake Detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2021, pp. 2185–2194. https://doi.org/10.1109/CVPR46437.2021.00222.
15. Zhao, T.; Xu, X.; Xu, M.; Ding, H.; Xiong, Y.; Xia, W. Learning Self-Consistency for Deepfake Detection. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021. https://doi.org/10.1109/ICCV48922.2021.01475.
16. Chen, S.; Yao, T.; Chen, Y.; Ding, S.; Li, J.; Ji, R. Local Relation Learning for Face Forgery Detection. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2021. https://doi.org/10.1609/aaai.v35i2.16193.
17. Fei, J.; Dai, Y.; Yu, P.; Shen, T.; Xia, Z.; Weng, J. Learning Second Order Local Anomaly for General Face Forgery Detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. https://doi.org/10.1109/CVPR52688.2022.01963.
18. Zhuang, W.; Chu, Q.; Tan, Z.; Liu, Q.; Yuan, H.; Miao, C.; Luo, Z.; Yu, N. UIA-ViT: Unsupervised inconsistency-aware method based on vision transformer for face forgery detection. In Proceedings of the European conference on computer vision. Springer, 2022, pp. 391–407. https://doi.org/10.1007/978-3-031-20065-6_23.
19. Huang, B.; Wang, Z.; Yang, J.; Ai, J.; Zou, Q.; Wang, Q.; Ye, D. Implicit Identity Driven Deepfake Face Swapping Detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. https://doi.org/10.1109/CVPR52729.2023.00436.
20. Li, L.; Bao, J.; Zhang, T.; Yang, H.; Chen, D.; Wen, F.; Guo, B. Face X-ray for More General Face Forgery Detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020. https://doi.org/10.1109/CVPR42600.2020.00505.
21. Cao, J.; Ma, C.; Yao, T.; Chen, S.; Ding, S.; Yang, X. End-to-End Reconstruction-Classification Learning for Face Forgery Detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 4113–4122. https://doi.org/10.1109/CVPR52688.2022.00408.
22. Chen, L.; Zhang, Y.; Song, Y.; Liu, L.; Wang, J. Self-Supervised Learning of Adversarial Example: Towards Good Generalizations for Deepfake Detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. https://doi.org/10.1109/CVPR52688.2022.01815.
23. Cheng, J.; Yan, Z.; Zhang, Y.; Luo, Y.; Wang, Z.; Li, C. Can We Leave Deepfake Data Behind in Training Deepfake Detector? In Proceedings of the Advances in Neural Information Processing Systems, 2024. https://doi.org/10.52202/079017-0691.

 doi:10.20944/preprints202604.1520.v1

24. Li, H.; Zhou, J.; Li, Y.; Wu, B.; Li, B.; Dong, J. FreqBlender: Enhancing DeepFake Detection by Blending Frequency Knowledge. In Proceedings of the Advances in Neural Information Processing Systems, 2024. https://doi.org/10.52202/079017-1429.
25. Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; Li, H. AltFreezing for More General Video Face Forgery Detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 4129–4138. https://doi.org/10.1109/CVPR52729.2023.00402.
26. Yan, Z.; Zhang, Y.; Fan, Y.; Wu, B. UCF: Uncovering Common Features for Generalizable Deepfake Detection. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023. https://doi.org/10.1109/ICCV51070.2023.02048.
27. Larue, N.; Vu, N.S.; Struc, V.; Peer, P.; Christophides, V. SeeABLE: Soft Discrepancies and Bounded Contrastive Learning for Exposing Deepfakes. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023. https://doi.org/10.1109/ICCV51070.2023.01921.
28. Dong, S.; Wang, J.; Ji, R.; Liang, J.; Fan, H.; Ge, Z. Implicit Identity Leakage: The Stumbling Block to Improving Deepfake Detection Generalization. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023. https://doi.org/10.1109/CVPR52729.2023.00389.
29. Yan, Z.; Luo, Y.; Lyu, S.; Liu, Q.; Wu, B. Transcending Forgery Specificity with Latent Space Augmentation for Generalizable Deepfake Detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024. https://doi.org/10.1109/CVPR52733.2024.00858.
30. Choi, J.; Kim, T.; Jeong, Y.; Baek, S.; Choi, J. Exploiting Style Latent Flows for Generalizing Deepfake Video Detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 1133–1143. https://doi.org/10.1109/CVPR52733.2024.00114.
31. Cui, X.; Li, Y.; Luo, A.; Zhou, J.; Dong, J. Forensics Adapter: Adapting CLIP for Generalizable Face Forgery Detection. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025. https://doi.org/10.1109/CVPR52734.2025.01789.
32. Yan, Z.; Yao, T.; Chen, S.; Zhao, Y.; Fu, X.; Zhu, J.; Luo, D.; Wang, C.; Ding, S.; Wu, Y.; et al. Df40: Toward next-generation deepfake detection. *Advances in Neural Information Processing Systems* **2024**, *37*, 29387–29434. https://doi.org/10.52202/079017-0925.
33. Li, Y.; Yang, X.; Sun, P.; Qi, H.; Lyu, S. Celeb-df: A large-scale challenging dataset for deepfake forensics. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 3207–3216. https://doi.org/10.1109/CVPR42600.2020.00327.
34. Li, Y.; Zhu, D.; Cui, X.; Lyu, S. Celeb-DF++: A Large-scale Challenging Video DeepFake Benchmark for Generalizable Forensics. *arXiv preprint arXiv:2507.18015* **2025**. https://doi.org/10.48550/arXiv.2507.18015.